Faculty of Science, Technology, Engineering and Mathematics
M140 Introducing statistics

Unit 6 Truancy

Contents

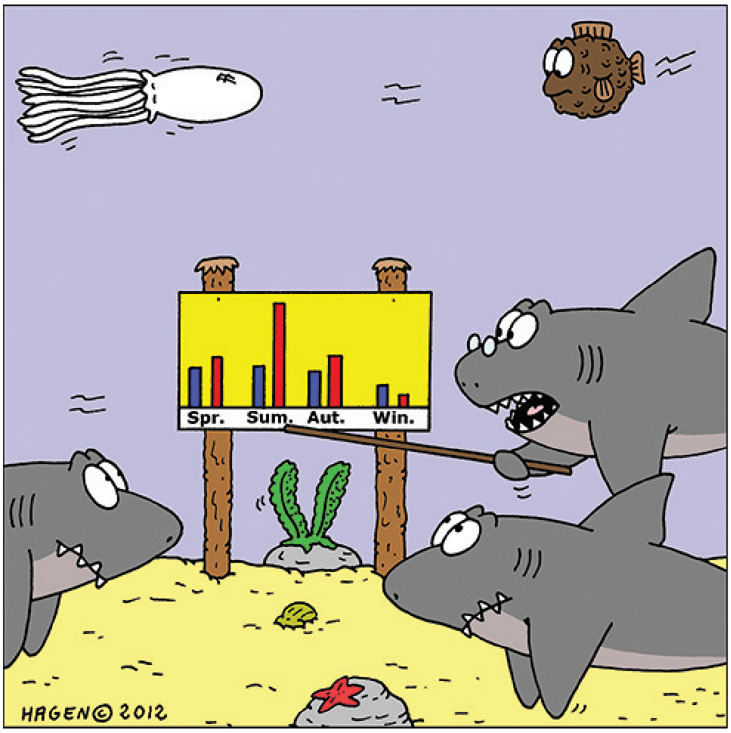
Introduction

In Unit 6 we address the following question:

How often do pupils truant?

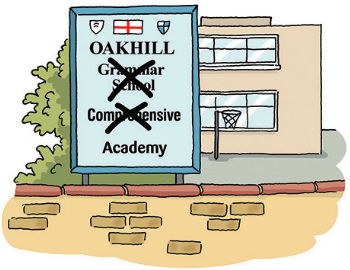
The topic of truancy has been chosen for this unit because it is an example that can be understood without technical background and which enables key statistical ideas to be illustrated and generalised. Statistics can enable informed discussion and decision-making to take place – decisions should be based on the careful analysis of reliable data. This is a module in statistics, not in education, but an aim in this unit and Units 7 to 9 is to demonstrate that statistics has an important role to play in the interpretation of data on education.

The emphasis in this unit is on using statistical techniques to reach conclusions; however, background factors which may account for the conclusions are also discussed. This will give you the opportunity to think about the issues involved, perhaps with particular reference to your own, or your friends', experiences.



Research shows that while the number of surfers is somewhat constant over the year, there is a sudden increase in casual bathers over summer.....

It is worth mentioning that approaches to education in the UK change regularly. This module was written in 2012–13, and if you are studying M140 some years later, the system may have a different structure, and different issues may be important. However, the same statistical techniques will be valid, and many of the questions will still be relevant.



Schools in England

The school system in England appears to be in a continual state of change. In the 1960s and 1970s, most grammar schools and secondary modern schools were replaced by comprehensive schools. Some City Technology Colleges started in the late 1980s and 1990s – they typically specialised in technology, science and mathematics, and often forged close links with businesses and industry.

More recently, many schools have been changing their status to become academies, thus gaining greater independence, particularly in respect of budget control. In 2010 there were only 200 academies, but the number had risen to 1635 by March 2012. Also, free schools were introduced in 2011 as an extension of the Academies Programme. These schools are funded by the taxpayer, but they are not controlled by the local authority and may be set up by parents, teachers, charities and businesses.

Unit 6, along with Units 7, 8 and 9, is a little more mathematical than earlier units. Do not worry if you find them difficult. The important thing is to ensure that you understand the results and how to apply them. In some cases, explanations are given of how results are derived. These are provided as you may find them interesting and they could increase your understanding. However, you will not be expected to reproduce these explanations in detail.

This unit contains six sections. In Section 1, we consider what we mean by the question *How often do pupils truant?*. How are we going to measure truancy? Which pupils are we talking about? We shall narrow the question down until we have a more specific one that we are able to answer. In Sections 2 and 3 we introduce the concept of *probability*, which is fundamental to the ideas of statistical inference. In Sections 4 and 5 we describe a strategy for drawing inferences about a population from a sample, and introduce a particular method, known as the *sign test*, that follows this strategy. Section 6 directs you to the Computer Book, where you will use Minitab to calculate certain probabilities and perform a sign test.

The unit introduces a lot of new concepts and ideas, some of which you may find difficult at first. However, you will find that these concepts occur many times in the remainder of the module, so you will have plenty of opportunity to become familiar with them.

Writing down numbers

In doing calculations, you have been encouraged to retain full accuracy in intermediate steps, perhaps writing down numbers *in full* to demonstrate that you are doing so. This is laborious and makes errors in copying down a number more likely. In the remaining units of this module, you need not do so. As far as you can, you should still use your calculator memory to store numbers to the calculator's accuracy, but you need not write down all numbers in full (usually about five significant figures will be enough, enabling you to focus on the most important part of a number).

1 Clarifying the question

In Section 1 we consider what is meant by the question

How often do pupils truant?

Random sampling was discussed in Subsection 1.2 of Unit 4.

Notice that this question refers implicitly to whole populations: for example, all schools in a particular area. We are usually not directly interested in how the children in one particular school behaved. However, it is often impossible, or at least not feasible, to collect data from the whole population. Instead we select a random sample of data. It might be a random sample of schools or of children. The sample is analysed by the methods we learned in earlier units, and we then need to decide how the results obtained from the sample apply to the whole population.

Statistical inference makes inferences about a population on the basis of data drawn from that population.

The above question about truancy may well have arisen from more general questions, such as:

Why do some pupils learn very little? Are we using good ways of teaching? Does the quality of my child's education depend on where I live?

However, these latter questions can only be tackled if they are first made more precise. Hence, rather than simply *posing* a question, we will often need to *clarify* it, and we may need to clarify it more than once as we learn more about the problem. In earlier units we used the modelling diagram shown in Figure 1 as a framework for how we explore and summarise batches of data.

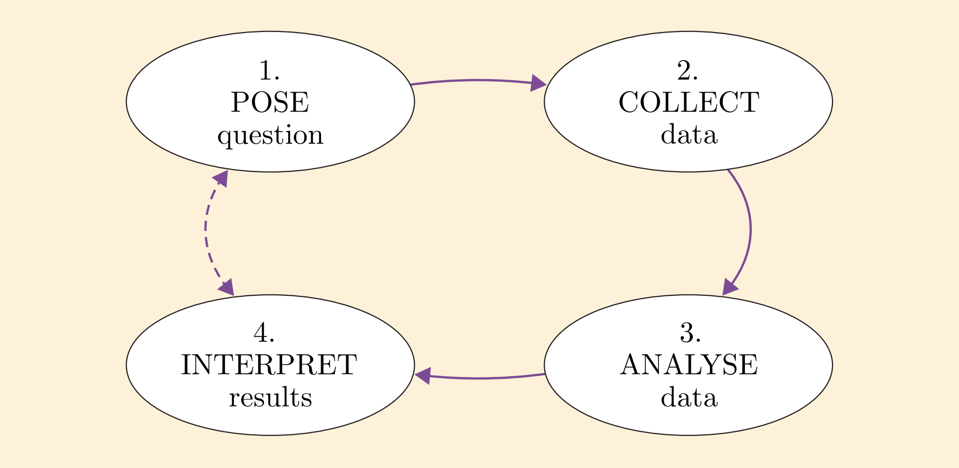


Figure 1 Modelling diagram

From now on we shall use a slightly modified form of this diagram in which

- the differing roles of populations and samples are identified
- the first box is changed from *pose question* to *clarify question*.

The modified modelling diagram is given in Figure 2.

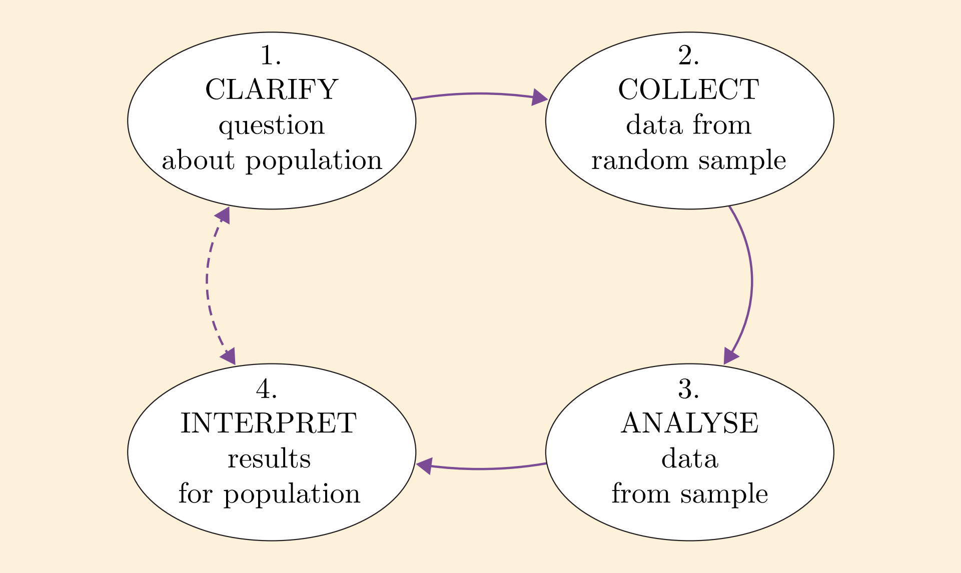
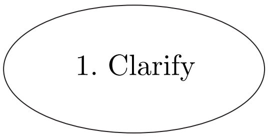


Figure 2 Modified modelling diagram

The processes involved in statistical inference are very important. While we are concentrating on them, we will be concerned particularly with stages 3 and 4 of the modelling diagram. However, this does not mean that we can ignore the other stages. It is always important to ensure that the question under discussion is defined carefully and precisely. Also, although in this unit we shall not be particularly concerned with *how* the data were collected, we do need to know the *form* of the data, as this affects the way the sample is analysed.

1.1 The question to be clarified

As we saw in earlier units, statistics is good at answering questions that require a numerical answer. However, the question for this unit is a very vague question. For example, we might be interested in how much particular children truant, or we might want to compare truancy at different schools.



First, suppose we were looking at particular children.

Activity 1 Factors affecting a child's truancy

Spend a few minutes thinking about what factors might affect how much a child plays truant. Then write down three factors that you think might be relevant.

Suppose a child psychologist is helping a particular child with a truancy problem. The psychologist would want to know the child's attendance record and factors about the child's circumstances that can influence truancy. The psychologist would then consider these factors and see if any pattern from the attendance record supported a given factor.

The same approach is followed if you move from considering individual truancy to truancy associated with different schools. We shall concentrate on looking at patterns with regard to schools, not individual children.

There are many different schools, and the amount of truancy will vary greatly. One of the interesting questions is whether different types of school have different amounts of truancy.

Activity 2 Factors affecting truancy in a school

Write down three factors that might affect the amount of truancy in a school.

Age of children is one of the most important factors in truancy figures. There is much less truancy at primary schools than at secondary schools. Young children are more likely to be taken to school by their parents, and also, since they are usually with the same class teacher all the time, truancy would be more easily noticed and could be followed up more quickly. We shall concentrate on secondary schools.

As you saw in the solution to Activity 2, there are still many factors that may affect truancy rate even after we have allowed for age to some extent by looking only at secondary schools. They include type of school, location of school and size of school, and there are also other factors, such as the attitude of the teachers, which are more difficult to measure. We shall look at several of these factors in the course of the unit, but we shall start with size of school.

A definition of ‘truancy rate’ is given in Subsection 1.2.

Is the truancy rate in large schools the same as the overall rate?

To simplify the problem further, and to try to consider a fairly homogeneous group of schools, we shall look at large schools in the East of England region and compare them with all schools in the East of England. It would not be sensible to compare large schools in the East of England with the national average truancy rate, as the East of England has a much lower truancy rate than, for example, Inner London. If we want to investigate the effect of size of school, we want the schools to be alike in *other* respects as much as possible, and location is one way of achieving this. So we shall ask the following question.

Do large secondary schools in the East of England have the same truancy rate as all secondary schools in the East of England?

Changes in pupil performance – up or down?

Pupils in the UK are steadily doing better when compared with their predecessors, while simultaneously doing worse when compared with the rest of the world. On the one hand, in 2011 the overall pass rate of A levels (qualifications typically used for university entrance in England, Wales and Northern Ireland) increased for the 29th consecutive year and more than a quarter of entries achieved the highest grade. Also in 2011, pass rates for Highers (qualifications typically used for university entrance in Scotland) hit a record high. On the other hand, Britain’s position in international education ratings dropped in most subjects after 2000.

The OECD (Organisation for Economic Co-operation and Development) publishes reports on economic and social factors in its member states, including school performance league tables. Between 2000 and 2006, the UK fell from 4th in the world to 14th in science, from 7th to 17th in literacy, and from 8th to 24th in mathematics. In part, no doubt, this reflects an increasing importance that other countries are placing on education. As Arne Duncan, the United States Secretary of Education, noted in a speech to UNESCO (United Nations Educational, Scientific and Cultural Organisation), ‘... in a knowledge economy, education is the new currency by which nations maintain economic competitiveness and global prosperity.’

1.2 Analysing the data

We have now decided on a specific question to investigate and we need to collect some data. This unit is mainly about analysing data that have already been collected, but it is worth spending a little time thinking about exactly what data should be collected. We can assume that we have a sampling frame consisting of all state-funded secondary schools in the East of England and the number of pupils they have. We can therefore pick out large schools, and we have arbitrarily defined these to be schools with 1000 or more pupils. We can then select a random sample of these schools. A sample size of 12 has been chosen.

A sampling frame was defined in Subsection 4.6 of Unit 4 as a list of all individuals in the target population.

We now want a single number to summarise the amount of truancy for each of the 12 schools. First we must consider what we mean by *truancy*. If a child skips school to go to the shops, then they are playing truant, while if they miss school because they are ill in bed, then they are not.

Activity 3 What is truancy?

Write down three reasons that a child might miss school – one that is definitely truancy, one that is definitely *not* truancy, and one that might or might not be truancy, depending on circumstances.

A school in Colorado photographed in 1915 during the season to harvest beet – only five pupils are at school while another thirty-five are absent because they are helping with the beet work.

A clear definition of truancy is needed if we are to gather truancy data for the different schools. The definition must take account of what data can be gathered, otherwise the definition may not be useful. In the next activity you are asked to think about how data related to truancy might be collected and used.

Activity 4 Measures of truancy

Think of two possible ways in which data on truancy in a particular school could be collected and truancy in the school measured. They should be feasible methods which will not occupy too much of the teachers' time.

When gathering data, precise definitions are needed. Hence the UK government collects data, not on truancy, but on 'unauthorised absence from school'. An unauthorised absence is absence without permission from a teacher or other authorised representative of the school. Records are kept of when permission for absence has been given (which would be retrospectively in the case of illness), so unauthorised absence is a well-defined, documented quantity. It is clearly closely related to truancy. Indeed, when the government publishes statistics on unauthorised absences from school, television and newspapers refer to them as **truancy rates**. We shall do the same.

Truancy rate

One of the statistics on schools that the government publishes is the **unauthorised absence rate**, and we will adopt this as our truancy rate.

- A *pupil's truancy rate* is the proportion of school half-days that the pupil was absent without authorisation.
- A *school's truancy rate* is the average truancy rate of its pupils.

You may well think this is not an ideal measure, and we shall return to this point in Subsection 5.3, but at present we shall concentrate on the analysis.

2. Collect

The module team has used data that were already published. The data give truancy rates over the first two terms of the 2010/2011 school year. The percentage truancy rates in a sample of 12 large schools in the East of England were as follows. (Note that from here on, 'school' means 'secondary school', unless otherwise qualified.)

Table 1 Truancy rates (%) in 12 large schools in the East of England

0.83	1.09	1.84	1.88	1.52	2.78	0.31	1.06	2.90	1.19	1.44	0.82

(Data source: Department for Education (2011) *Pupil absence in schools in England, autumn term 2010 and spring term 2011*)

The median truancy rate for all secondary schools in the East of England during the first two terms of 2010/2011 is known to be 0.98%.

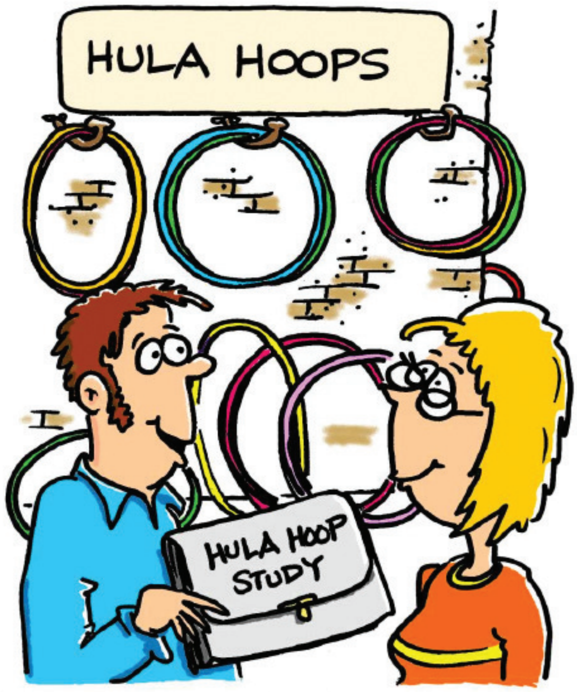
This figure of 0.98% is a population median; see Subsection 3.2 of Unit 4.

We can now move on to stage 3 and then stage 4 of the modelling diagram. We can analyse our sample and then try to interpret our results in terms of the whole population. This last step is known as *statistical inference*.

Statistical inference

This is a process of inferring back from the sample to the population. Somewhat paradoxically, the best approach to making inferences about populations from a sample is to consider what samples are likely to be obtained from a population.

To carry out statistical inference successfully, it is important that we have selected a *random* sample.



'The population of hula hoops tells you what kind of hula hoops will be in your sample. But the sample of hula hoops is used to infer things about the population. Isn't that circular logic?'

Activity 5 Median truancy rate of sample of 12 large schools

(a) Find the median truancy rate of the sample of 12 schools.

(b) Can you infer back from the sample to the population of all the East of England schools? In particular, is it possible to infer that the population median of all large schools in the East of England is, or is not, equal to 0.98%?

We do not expect you to give a definitive answer to part (b). Just think a little about the questions asked.

The sample median rate of 1.32 is larger than 0.98 but not by an enormous amount. If we had chosen a different random sample, the median would probably have been different. Here is another random sample of size 12 of truancy rates in large schools in the East of England. This time the values have been written in ascending order:

3. Analyse

0.59, 0.67, 0.68, 0.94, 1.35, 1.38, 1.48, 1.54, 1.61, 1.86, 1.89, 1.99.

In this case the sample median is $\frac{1}{2}(1.38 + 1.48) = 1.43$. This is also bigger than 0.98 and a little larger than the sample median found in Activity 5.

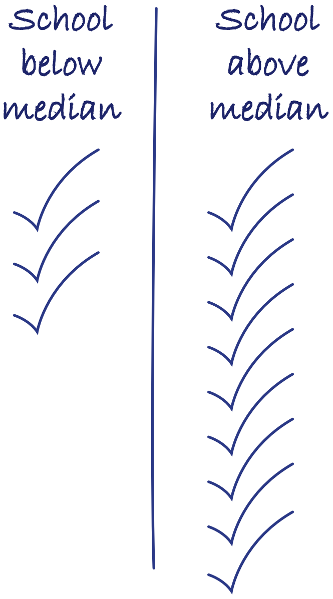
We should not infer back simply by looking at the sample and guessing. We need a systematic method, and the development of such a method is the main purpose of this unit.

4. Interpret

We shall look at the data of Table 1 in a different way. Instead of considering the sample median, we shall take each value in turn and record whether it is larger or smaller than 0.98. The first value of 0.83 is below 0.98, so we shall record that as a negative difference, or [−]. The value 1.09 lies above 0.98, so we record it as [+]. Working through the whole sample, we obtain:

0.83	1.09	1.84	1.88	1.52	2.78	0.31	1.06	2.90	1.19	1.44	0.82
[−]	[+]	[+]	[+]	[+]	[+]	[−]	[+]	[+]	[+]	[+]	[−]

Altogether, we have nine values above 0.98 and three values below it.



In Section 4, we are going to assume that the median percentage truancy rate for large schools in the East of England is indeed 0.98. We are then going to try to answer the question:

> *How likely is it that a random sample of size 12 would contain nine values above and three values below the assumed population median?*

To answer this question, we need a way of making an informed judgement.

If the median truancy rate for large schools were 0.98, it would not be surprising if there were six values above 0.98 and six values below it. This is the result that seems most likely, since the definition of a population median is that half the population values lie above it and half lie below it. We should expect a representative sample to be similar to the population. Also, after our work with samples in Unit 4, we should not be very surprised to find seven values above the median and five below it (or the other way round).

However, it would seem surprising if all 12 values were either above or below the population median. So if all 12 were greater than 0.98 (the assumed population median), we would suspect that the population median for large schools was larger than 0.98.

This suggests that some events are more likely to occur than others. The next activity suggests you think further along these lines.

Activity 6 Which events are more likely to happen?

Look at the following list of events and rank them in order of likeliness to occur, starting with the one you think most likely.

A. Your new colleague at work has the same birthday as you.

B. Two out of a group of ten people have the same birthday.

C. The sun will rise tomorrow.

D. The sun will shine tomorrow.

E. You will win the jackpot in the National Lottery next week.

F. You toss a coin and it lands tails up.

G. You throw a die and it shows a six. (Remember 'die' is the singular form of 'dice'.)

H. A member of a hockey team is 150 years old.

Activity 6 has demonstrated that some events are certain to occur, others are impossible, and the likelihoods of occurrence of others are somewhere between these two extremes. In the next two sections, we shall introduce the concept of probability, which is a method of associating a numerical measure with the likelihood of occurrence of a particular event.

We shall then be able to use this concept to calculate how likely we are to observe nine values above and three values below the population median in a random sample of size 12. Then we shall be in a better position to use the sample to make inferences about the population and decide whether or not the median truancy rate of the population of large secondary schools in the East of England during the first two terms of 2010/2011 is equal to 0.98%.

Exercises on Section 1

Exercise 1 Your view on the likely occurrence of more events

Look at the following list of events and rank them in order of likeliness to occur, starting with the one you think most likely.

A. A husband and wife find they were born on the same day of the week.

B. England will win the next football World Cup.

C. A Brazilian team will win the next football European Cup.

D. Death and taxes.

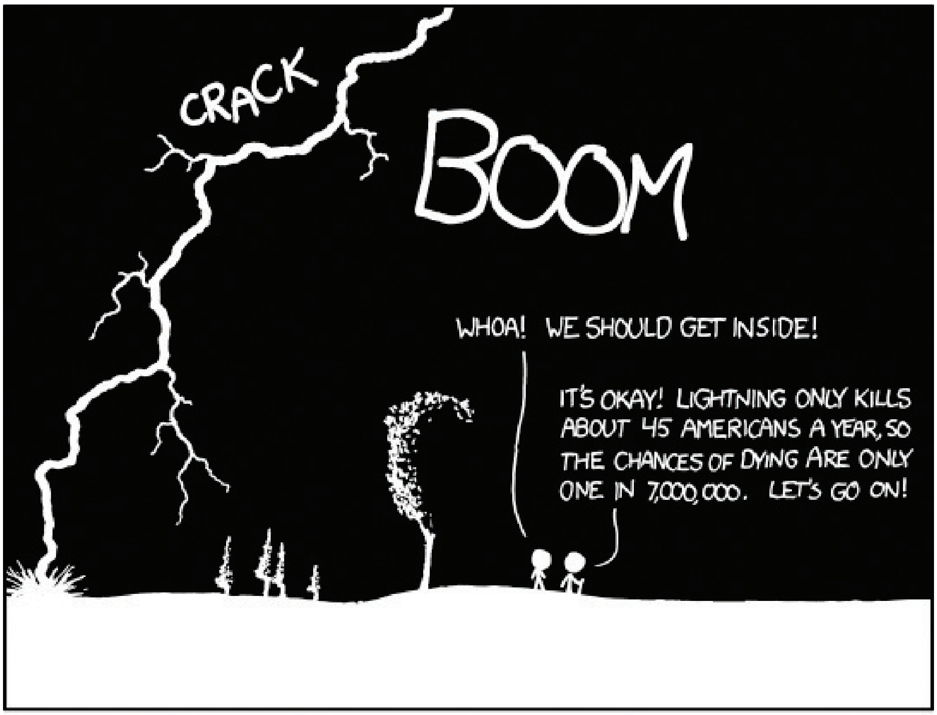
E. A mother's first pregnancy results in twins.

F. You get exactly three heads when you toss a coin five times.

G. A chicken egg has two yolks.

H. It snows in London on Christmas day.

I. You will be struck by lightning next year.


CRACK
BOOM
WHOA! WE SHOULD GET INSIDE!
IT'S OKAY! LIGHTNING ONLY KILLS ABOUT 45 AMERICANS A YEAR, SO THE CHANCES OF DYING ARE ONLY ONE IN 7,000,000. LET'S GO ON!

THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

2 Probability

At the end of Section 1, we established it would be useful to associate a number with an event so that it provided a measure of the likelihood of that event. This number is called the **probability** of occurrence of the event. The concept of probability plays such an important role in statistics that we are going to spend two sections investigating its properties, before returning to our specific question of truancy.

'How do you want it – the crystal mumbo-jumbo or statistical probability?'

2.1 Measuring chance

The easiest way of thinking about probability is to equate it to proportion: the *probability* that a particular event will happen is the *proportion* of the time that it is expected to happen. When we toss a fair (unbiased) coin, for example, there is a fifty-fifty chance that the coin will land 'heads' because half the time it should land 'heads' and half the time it should land 'tails'. That is, the proportion of time that the outcome should be heads is $\frac{1}{2}$, so the *probability* that the outcome will be heads is $\frac{1}{2}$.

(In practice, of course, you can only toss a coin a finite number of times, and it is very unlikely to land 'heads' exactly half the time. For example, if you toss it 600 times, then there is little chance that it will land 'heads' exactly 300 times. However, if you toss a coin an enormous number of times, the proportion of 'heads' should be very close to $\frac{1}{2}$.)

Similarly, if a die is fair, then each of its six sides is equally likely to be the outcome when it is rolled. Thus, for example, the proportion of rolls that should result in a 4 is $\frac{1}{6}$, so the probability of rolling a 4 is $\frac{1}{6}$.

Probability = Proportion

See Subsection 1.2 of Unit 4.

You met the ideas of random selection and random sampling in Unit 4. With random sampling, each member of the population is equally likely to be included in the sample. In particular, if a person or item is picked at random from a population, each member of the population is equally likely to be the one that is picked. We shall use these ideas to begin our investigation of probability.

Example 1 Student probabilities

Table 2 gives the breakdown of male and female students in a small university, by their faculty. (A student belongs to exactly one faculty.)

Table 2 Student population of a small university

	Science	Arts	Law	Medicine	Total
Female	964	1532	87	206	2789
Male	1638	1039	247	369	3293
Total	2602	2571	334	575	6082

Now suppose that a student is selected at random from the university. Using the notion that probability equates to proportion, various probabilities relating to the student can be calculated. For example, the table gives the number of students who are female (2789) and the total number of students (6082). Hence we can determine the probability that the selected student is female by calculating the proportion of students who are female, as follows.

The probability of selecting a female student is

$$\frac{\text{number of female students}}{\text{total number of students}} = \frac{2789}{6082} \simeq 0.459.$$

Similarly, the probability of selecting an arts student is

$$\frac{\text{number of arts students}}{\text{total number of students}} = \frac{2571}{6082} \simeq 0.423,$$

and the probability of selecting a male science student is

$$\frac{\text{number of male science students}}{\text{total number of students}} = \frac{1638}{6082} \simeq 0.269.$$

You have now covered the material related to Screencast 1 for Unit 6 (see the M140 website).

Activity 7 Student probabilities

Giving your answers to three decimal places, what is the probability that a student chosen at random from the university population in Table 2 is:

(a) a male student?

(b) a law student?

(c) a female medical student?

Whenever you perform a calculation, you should look to see if your answer seems sensible – that way, you can sometimes spot large calculation errors. In the last activity you would expect your probability for (a) to be large because a lot of the students are male, while the probabilities for (b) and (c) should be small because there are comparatively few law students or female medical students.

Activity 8 More probabilities

Suppose we have a population of men and women and select just one person at random. For each of the following cases, give the probability that a woman is selected.

(a) The population of size 10 consists of 5 men and 5 women.

(b) The population of size 10 consists of 1 man and 9 women.

(c) The population of size 10 consists of 9 men and 1 woman.

(d) The population of size 10 consists of 10 men and 0 women.

(e) The population of size 100 consists of 99 men and 1 woman.

(f) The population of size 100 consists of 0 men and 100 women.

In part (d) of Activity 8, the probability of selecting a woman was 0. This particular population contained no women, so it was impossible to select a woman. Similarly in part (f), the probability of selecting a woman was 1, and in that situation, the population contained no men, so a woman was certain to be selected.

If you look at the other probabilities we have calculated, you will see that they are all between 0 and 1. This is because they have all been calculated as the proportion of a population who possess a certain property – for example, being a science student or being a woman. The closer the probability is to 1, the more likely the event is; the closer the probability is to 0, the rarer the event is. If you ever calculate the probability of an event and find it is either greater than 1 or a negative number, then you know you have made a mistake somewhere.

This discussion can be summarised by the following three important properties of probability:

- If an event is impossible, then its probability is 0.
- If an event is certain, then its probability is 1.
- Any event which is uncertain but not impossible has a probability that lies between 0 and 1.

At this point it is convenient to introduce some notation which cuts down the number of words we have to write. In the last activity you looked at the probability of the event that a person selected at random from a population is a woman. If we let W stand for the event *a woman is selected*, then we shall write $P(W)$ for the probability that event W occurs, or the probability that a woman is selected. So

Both P and Pr are commonly used to mean ‘probability’ – we shall use P.

$$P(W) = \frac{\text{number of women in population}}{\text{total number in population}}.$$

Similarly, we might let L stand for the event that a student selected at random is a law student. Then

$$P(L) = \frac{\text{number of law students in population}}{\text{total number in population}}.$$

This can be extended to a more general situation.

Let E stand for the event of selecting a person or object with some particular property from a population, using random sampling. Then the **probability** of E is given by

$$P(E) = \frac{\text{number in population with particular property}}{\text{total number in population}}.$$

2.2 Adding probabilities

Probability gives a way of measuring how likely an event is to occur in random sampling. In the last subsection you learned that a probability is always greater than or equal to 0 and always less than or equal to 1. The following example and activity use data on truancy to help you become more familiar with the idea of probability. The same data is then used to explore another property of probability.

Example 2 Truancy in two schools

Table 3 shows some invented data on truancy in two schools, A and B, that contained 200 and 100 pupils, respectively.

We have invented the data so that the numbers are small and the calculations are easy to check.

Table 3 Truancy numbers in two schools

Number of days absent	Number of pupils School A	School B	Total numbers
0 to 4	108	42	150
5 to 9	60	30	90
10 to 19	26	19	45
20 or more	6	9	15
Total	200	100	300

We shall use the table to answer the following questions.

1. If a child is selected at random from these two schools, what is the probability that this child was absent through truancy for fewer than 5 days?
2. If a child is selected at random from these two schools, what is the probability that this child is at School A and was absent through truancy for fewer than 5 days?
3. If a child is selected at random *from School A*, what is the probability that this child was absent through truancy for fewer than 5 days?

Let T stand for the event that a child selected at random was absent through truancy for fewer than 5 days, and let A stand for the event that the child attends School A.

(a) Here the probability is $P(T)$. Now

$$P(T) = \frac{\text{total number of children absent for} < 5 \text{ days}}{\text{total number of children}}$$
$$= \frac{150}{300} = 0.5.$$

So there is a probability of 0.5 that a child picked at random from these two schools was absent through truancy for fewer than 5 days.

The symbol $<$ means 'less than'. Similarly, $>$ means 'greater than', $\leq$ means 'less than or equal to' and $\geq$ means 'greater than or equal to'.

(b) Here the probability is that both events T and A occur. This is $P(T \text{ and } A)$, which is an extension of our notation for the probability of an event. (It means the probability that both T and A occur. In this case, the event 'T

and A' occurs if a child is absent through truancy for fewer than 5 days and also attends School A.) From Table 3, we see that 108 children attended School A and were absent through truancy for fewer than 5 days. So

$$\begin{aligned} P(T \text{ and } A) &= \frac{\text{total number of children satisfying both } T \text{ and } A}{\text{total number of children}} \\ &= \frac{108}{300} = 0.36. \end{aligned}$$

So the probability that a child attends School A and is absent through truancy for fewer than 5 days is 0.36.

(c) You have to be careful here to appreciate what probability is required and how it differs from part (b). Probability is giving us precise answers, so we have to be careful that we ask precise questions and that the questions are the ones we want to answer. Here we assume that the child is selected at random from those at School A, so School A provides the total population. Thus

$$\begin{aligned} &P(\text{child from School A is absent for} < 5 \text{ days}) \\ &\quad = \frac{\text{total number of children at School A absent for} < 5 \text{ days}}{\text{total number of children at School A}} \\ &\quad = \frac{108}{200} = 0.54. \end{aligned}$$

So the required probability is 0.54.

Activity 9 Truancy in two schools

(a) Find the probability of each of the following events, using the data in Table 3.

- A child chosen at random from these two schools is absent through truancy for between 5 and 9 days.
- A child chosen at random from these two schools is absent through truancy for between 5 and 9 days and attends School B.
- A child chosen at random from those attending School A is absent through truancy for between 10 and 19 days.
- A child chosen at random from these two schools is absent through truancy for 10 or more days.

(b) For the first two events in part (a), define letters to denote the appropriate events as in Example 2. Use these letters to write in $P(E)$ notation the probabilities you calculated.

We next consider when probabilities can be added together. In Activity 9(a) you calculated the probability that a child selected at random from the two schools is absent for 10 or more days. This probability is 0.20. Now absence for 10 or more days means either absence for 10 to 19 days or absence for 20 or more days; in the context of the question, it does not matter which of these two events actually occurs.

We can easily find the probability of occurrence of each of these events:

$$\begin{aligned} P(\text{child absent for 10 to 19 days}) &= \frac{45}{300} = 0.15, \\ P(\text{child absent for 20 or more days}) &= \frac{15}{300} = 0.05. \end{aligned}$$

The sum of these two probabilities is $0.15 + 0.05 = 0.20$, which is the probability that a child selected at random is absent for 10 or more days. So we have

$$\begin{aligned}&P(\text{child absent for} \geq \text{10 days})\\&\quad = P(\text{child absent for 10 to 19 days}) + P(\text{child absent for} \geq \text{20 days}).\end{aligned}$$

This is clearly not a coincidence, since

$$\begin{aligned}&P(\text{child absent for} \geq \text{10 days})\\&\quad = \frac{\text{total number of children absent for} \geq \text{10 days}}{\text{total number of children}}\\&\quad = \frac{\text{number absent for 10 to 19 days} + \text{number absent for} \geq \text{20 days}}{\text{total number of children}}\\&\quad = \frac{\text{number absent for 10 to 19 days}}{\text{total number of children}} + \frac{\text{number absent for} \geq \text{20 days}}{\text{total number of children}}\\&\quad = P(\text{child absent for 10 to 19 days}) + P(\text{child absent for} \geq \text{20 days}).\end{aligned}$$

This example was about two events, either of which might occur. However, it is impossible that both events occur at the same time. A child cannot both be absent for 10 to 19 days and for 20 or more days. Any two events with this property are called *mutually exclusive*. (The two events ‘absent for 10 or more days’ and ‘absent for 20 or more days’ are not mutually exclusive. If a child were absent for 21 days, both of these events would occur.)

Two events are said to be **mutually exclusive** if they cannot occur at the same time. More generally, any number of events are said to be mutually exclusive if no two of them can occur at the same time.

Activity 10 Exclusive events?

Which of the following pairs or sets of events are mutually exclusive?

(a) A person’s blood type is measured and the events are: (i) it is blood type A, and (ii) it is blood type O.

(b) A person is described and the events are: (i) the person has black hair, and (ii) the person has blue eyes.

(c) Some dice are rolled and the events are: (i) 3 dice are rolled, (ii) each die gives the same number, and (iii) the sum of the numbers on the dice is 8.

(d) A person’s blood type is measured and the events are: (i) it is blood type A, (ii) it is blood type O, and (iii) the person’s blood is rhesus positive.

Suppose A and B are two mutually exclusive events. Since they are mutually exclusive events, both of them cannot occur at the same time. We shall denote the probability that one of A and B occurs by $P(A \text{ or } B)$. This probability is given by the addition rule for mutually exclusive events.

Addition rule for mutually exclusive events (the ‘or’ linkage)

For any two mutually exclusive events A and B,

$$P(A \text{ or } B) = P(A) + P(B).$$

This rule extends to more than two events when they are all mutually exclusive. For example, if A, B and C are mutually exclusive events, then

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C).$$

Activity 11 Adding two probabilities

This activity relates to the data in Table 3 in Example 2, reproduced below for convenience.

Table 4 Truancy numbers in two schools

Number of	Number of pupils		Total numbers
days absent	School A	School B	
0 to 4	108	42	150
5 to 9	60	30	90
10 to 19	26	19	45
20 or more	6	9	15
Total	200	100	300

(a) Suppose a child is selected at random from School A. Find the probabilities that this child is absent through truancy for

- 0 to 4 days
- 5 to 9 days
- 0 to 9 days.

(b) Verify that the addition rule for mutually exclusive events holds in this case.

There is a particular application of the addition rule which gives us a useful result about probabilities. Suppose E is some event that may or may not occur at a certain trial. (In statistics, a sequence of observations or a sequence of tests is often called a set of *trials*.) For example, the trial might be the selection of one child from the population of the two schools, and E might be the event that the child is absent through truancy for 0 to 4 days. Let F be the event that E does *not* occur. Then F is the event that a child is absent for 5 or more days. The events E and F are called **complementary**, because only one of them can happen and together they *complete* the list of possibilities.

As one of the events E or F must occur, $P(E \text{ or } F) = 1$. (Think about this. Is it possible that neither E nor F occurs?)

Hence, by the addition rule,

$$P(E \text{ or } F) = P(E) + P(F) = 1.$$

So $P(F) = 1 - P(E)$. We express this rule in words, as follows.

Probability rule for complementary events (the ‘not’ linkage)

For any event E,

$$P(E \text{ does not occur}) = 1 - P(E \text{ does occur}).$$

You have now covered the material related to Screencast 2 for Unit 6 (see the M140 website).

Activity 12 A small university revisited

Table 2 (reproduced below as Table 5) gave the breakdown of male and female students in a small university, by their faculty.

Table 5 Student population of a small university

	Science	Arts	Law	Medicine	Total
Female	964	1532	87	206	2789
Male	1638	1039	247	369	3293
Total	2602	2571	334	575	6082

Suppose that a student is selected at random.

(a) Find the probability that the student is male. Hence determine the probability that the student is female.

(b) What is the probability that the student is not a medical student?

Activity 13 Occasional truant, or not

In relation to the data in Table 3 (reproduced in Table 4), let G be the event that a child selected at random from the two schools is absent through truancy for 5 to 9 days.

(a) Describe the event 'G does not occur'.

(b) Use the above rule to evaluate: $P(G$ does not occur).

2.3 Multiplying probabilities

We have seen that probabilities are added when we have the 'or' linkage, and want P(A *or* B). We next consider how to determine probabilities when we have an 'and' linkage, and want P(A *and* B). We use the notion that P(A and B) is the proportion of the time that A and B both happen.

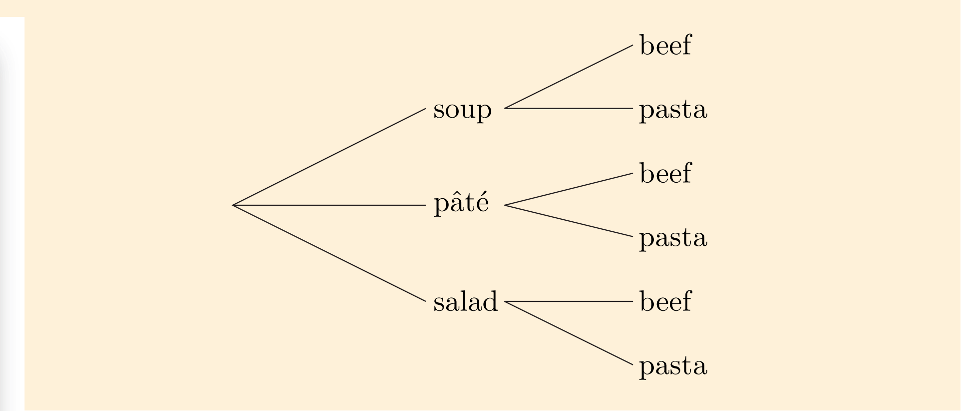
Example 3 Two-course lunch

A restaurant offers a two-course set lunch. There are three choices for the first course – soup, pâté or salad – and two choices for the second course – beef or pasta. The different meal-combinations are shown in Figure 3.

Two-course set menu

Soup
Pâté
Salad

Beef
Pasta



gure 3 Tree diagram for a two-course lunch

The diagram in Figure 3 is referred to as a **tree**. Starting at the left of the figure, we can follow one of three lines – **branches** – to choose a first course (soup, pâté or salad). From each first course we can follow one of two lines – **sub-branches** – to choose the second course (beef or pasta). Thus there are $3 \times 2 = 6$ different paths we can follow, corresponding to the six possible meal combinations: soup–beef, soup–pasta, pâté–beef, pâté–pasta, salad–beef and salad–pasta.

Suppose, now, that we choose a first course at random and also choose the second course at random. Then each of these six possibilities is equally likely. Thus the proportion of time we choose, say, salad followed by beef would be one-sixth, so

$$P(\text{salad and beef combination}) = \frac{1}{6}.$$

Notice that there is a choice of three first courses, so if the choice is made at random,

$$P(\text{salad for first course}) = \frac{1}{3}.$$

And, as there are two choices for the second course,

$$P(\text{beef for second course}) = \frac{1}{2}.$$

Consequently, in this example

$$P(\text{salad and beef combination}) = P(\text{salad}) \times P(\text{beef}).$$

Extending Example 3 is helpful, so suppose that there are four choices for the first course – soup, salad, pâté and prawns – and five choices for the second course – beef, chicken, fish, pasta and quiche. Following similar reasoning to Example 3, there are $4 \times 5 = 20$ different meal combinations.

Activity 14 Two-course meal

(a) Draw a diagram similar to Figure 3 to show the 20 different meal combinations.

(b) Suppose the soup and salad are both vegetarian options for the first course, and pasta and quiche are vegetarian options for the second course. Write down the different combinations of courses that a vegetarian could have.

(c) If a first course and second course are selected at random, what is the probability that the combination is vegetarian?

(d) Suppose a first course and second course are selected at random.

- What is the probability that the first course is vegetarian?
- What is the probability that the second course is vegetarian?

Check that the multiplication of these answers gives the probability you found in (c).

In Activity 14 you found that

$$\begin{aligned}&P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course})\\&\quad = P(\text{vegetarian first course}) \times P(\text{vegetarian second course}).\end{aligned}$$

The reason is as follows:

$$\begin{aligned}&P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course})\\&\quad = \frac{\text{number of vegetarian meal combinations}}{\text{total number of meal combinations}}\\&\quad = \frac{\text{number of vegetarian first courses} \times \text{number of vegetarian second courses}}{\text{total number of first courses} \times \text{total number of second courses}}\\&\quad = \frac{\text{number of vegetarian first courses}}{\text{total number of first courses}} \times \frac{\text{number of vegetarian second courses}}{\text{total number of second courses}}\\&\quad = P(\text{vegetarian first course}) \times P(\text{vegetarian second course}).\end{aligned}$$

To extend this result, the following activity concerns three-course meals.

Activity 15 Three-course meal

Suppose the menu for a three-course meal has the same choices of first course and second course as in Activity 14 and it additionally has a third course with three options: pie, crumble or ice cream.

(a) How many different three-course meal combinations are there?

(b) Suppose the three courses are each selected at random.

- What is the probability that the third course would be good with custard (i.e. pie or crumble)?
- What is the probability that the first two courses are vegetarian and the third course is good with custard?

(c) When the three courses are selected at random, check that

$$\begin{aligned}&P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course}\\&\qquad \textit{and} \text{ third course is good with custard})\\&\quad = P(\text{vegetarian first course}) \times P(\text{vegetarian second course})\\&\qquad \times P(\text{third course is good with custard}).\end{aligned}$$

What's the probability of this meal?

We now formulate in general terms the rule we have been using which enables us to multiply probabilities in appropriate circumstances. To do this we need the concept of 'statistical independence'.

Two events are said to be **statistically independent** if the occurrence of one has no effect on the likelihood of occurrence of the other.

In the last activity, for example, we assumed that each course was chosen at random, so the choice made for one course had no influence on the choice made for any other course. Thus, for example, the two events ‘choosing beef for the second course’ and ‘choosing crumble for the third course’ are statistically independent.

Activity 16 Independent events?

Which of the following pairs or sets of events are independent?

(a) Picking a person at random where the events are: (i) the person is taller than average; (ii) the person is heavier than average.

(b) Picking two people at random where the events are: (i) the first person is taller than average; (ii) the second person is heavier than average.

(c) Picking a person at random where the events are: (i) the person is taller than average; (ii) the person was born on a Tuesday.

(d) Tossing two coins where the events are: (i) the first coin lands ‘heads’; (ii) the second coin lands ‘heads’; (iii) both coins give the same outcome.

(e) Drawing a card from an ordinary deck of playing cards where the events are: (i) the card is an ace; (ii) the card is a diamond.

Let A and B be two statistically independent events, and let ‘A and B’ denote the event that both A and B occur. Then the probability that A and B both occur is given by the following rule.

Multiplication rule for statistically independent events (the ‘and’ linkage)

If A and B are statistically independent events, then the probability that A and B both occur is given by

$$P(A \text{ and } B) = P(A) \times P(B).$$

This rule extends to more than two events. For example, if A, B and C are statistically independent events, then the probability that they all occur is given by

$$P(A \text{ and } B \text{ and } C) = P(A) \times P(B) \times P(C).$$

You have now covered the material related to Screencast 3 for Unit 6 (see the M140 website).

Earlier in this section, we determined probabilities when an individual was picked at random from a population. In practice, often we randomly sample a number of individuals or a number of items from a population. We shall use the *addition rule* and *multiplication rule* to look at probabilities for samples that do not consist of just one individual but of several. First we will look at probabilities for samples of size 2, before seeing how our results can be generalised to larger samples.

When we introduced random samples in Unit 4, we saw how they could be selected using random number tables. We selected a random number to give us the first member of the sample, then a second, and so on. If the same random number appeared a second time, we rejected it, because the same individual cannot appear more than once in a random sample. For the moment, we shall relax this restriction, because it makes the probabilities easier to find. Thus, if an individual or item is selected for a random sample, it can be picked again at a second selection. Later in this section, we shall return to proper random sampling and show that in practice for large populations, it makes hardly any difference to the probabilities.

We shall introduce the method of finding probabilities for a sample of size 2 by means of an example.

Example 4 Sample of two

Suppose we have a population of size 10 which contains three women and seven men. We shall call the three women Ashia, Brenda, and Clare, and the seven men David, Ejike, Frank, Gavin, Harry, Ian and Joe, and refer to them all by their initials. Suppose we select a sample of size 2 from this population, remembering that we may select the same person twice.

How many samples of size 2 are there altogether? We must differentiate between the person selected the first time and the person selected the second time. We might pick Joe first and Clare second; we could write this as (J, C). Alternatively, we might pick Clare first and Joe second; this is (C, J). Other samples include (E, F) and (A, A), as Ashia might be selected twice. There are 10 possibilities for the first member of the sample and 10 possibilities for the second member. Altogether there are $10 \times 10 = 100$ possible samples of size 2.

Activity 17 Sample of two women

(a) For the population introduced in Example 4, write down all the samples of size 2 that contain two women, using their initials A, B, C. How many are there?

(b) How could you have found this number without writing down all the samples?

So using the results from Example 4 and Activity 17, we can now write down the probability that our sample of size 2 from this population of ten people contains two women.

Equating probability to proportion, we can write

$$\begin{aligned}&P(\text{sample of size 2 contains two women})\\&\quad= \frac{\text{number of samples of size 2 containing two women}}{\text{total number of samples of size 2}}\\&\quad= \frac{9}{100} = 0.09.\end{aligned}$$

We might also obtain this probability by applying the multiplication rule. Now the probability that the first selection is a woman is $3/10 = 0.3$, as is the probability that the second selection is a woman. So to obtain the probability that both selections are women, we multiply the probabilities together. We shall introduce some notation to describe this result. Let W denote the event that at any selection a woman is chosen, and let $2W$ denote the event that two women are selected in a sample of size 2. Then

$$P(2W) = P(W) \times P(W) = 0.3 \times 0.3 = 0.09.$$

We shall continue to sample from the population of Ashia, Brenda, Clare, . . . , Joe.

Activity 18 Sample of two men

Suppose M denotes the event that at any selection a man is chosen, and $2M$ denotes the event that two men are selected in a sample of size 2. Use the multiplication rule to find $P(2M)$.

We can use both this multiplication rule and the addition rule from Subsection 2.2 to find the probability that a sample of size 2 contains one woman and one man. To obtain such a sample, we either select a woman first and a man second, or a man first and a woman second. These possibilities are mutually exclusive events, so we add the probabilities:

$$\begin{aligned}&P(\text{sample of size 2 contains one woman and one man})\\&\quad= P(\text{a woman is selected first and a man is selected second})\\&\qquad+ P(\text{a man is selected first and a woman is selected second}).\end{aligned}$$

Now we know that at any selection, $P(W) = 0.3$ and $P(M) = 0.7$. So, just as with a sample of two women, we multiply the probabilities together to obtain each of the probabilities on the right-hand side. Writing $P(1W)$ for the probability that a sample of size 2 contains one woman (and therefore one man), we obtain

$$\begin{aligned}P(1W) &= P(W) \times P(M) + P(M) \times P(W)\\&= 0.3 \times 0.7 + 0.7 \times 0.3\\&= 0.21 + 0.21 = 0.42.\end{aligned}$$

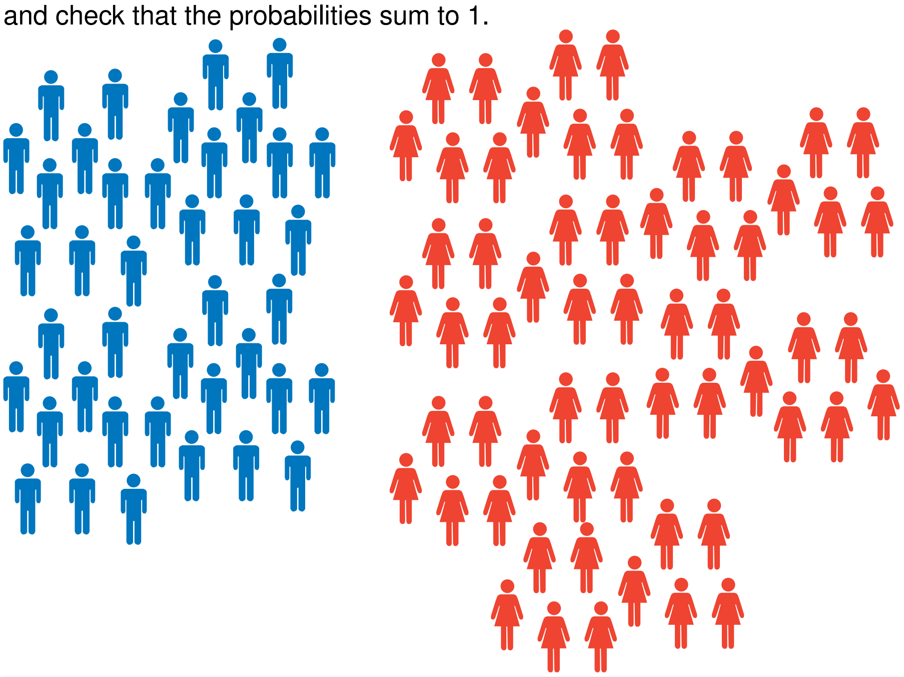
We can now check our results by using the fact that our sample is certain to contain either two women or two men, or one woman and one man; that is,

$$P(2W) + P(2M) + P(1W) = 1.$$

Substituting the calculated values, we obtain $0.09 + 0.49 + 0.42 = 1$, as required.

Activity 19 Sampling from 100 people

A population contains 40 men and 60 women. A sample of size 2 is drawn in which the first person selected is also available for selection the second time. Find the probabilities for the different possible numbers of men in the sample, and check that the probabilities sum to 1.



Earlier, we decided to introduce the calculation of probabilities for samples of size 2 by allowing the same individual to be selected twice. Now we shall consider probabilities for random samples of size 2, when any individual can appear only once in a sample. Let us start by looking again at the population of Example 4 (Ashia, Brenda, Clare, . . . , Joe), which consists of three women and seven men. We can select the first member of the sample in 10 ways. Then only nine individuals remain, so we can select the second in 9 ways. Hence the total number of samples of size 2 is $10 \times 9 = 90$.

Activity 20 Sampling two different women

How many random samples of size 2 contain two different women? Remember that Ashia first and Brenda second is a different sample from Brenda first and Ashia second, and that now the same person cannot be selected twice.

Using the result of Activity 20, we can find

$$P(\text{random sample of size 2 contains two women})$$
$$= P(2W) = \frac{6}{90} = \frac{1}{15} = 0.067 \quad \text{(rounded to three decimal places).}$$

Similarly, the probability of there being two men in the sample is

$$\frac{7 \times 6}{90} = \frac{42}{90} = \frac{7}{15} \simeq 0.467.$$

So, by the addition rule and the probability rule for complementary events, the probability of one man and one woman is

$$1 - \left(\frac{1}{15} + \frac{7}{15}\right) = \frac{7}{15} \simeq 0.467.$$

These probabilities – 0.067, 0.467, 0.467 – are different from those we found before when the same individual could be selected twice: 0.09, 0.49, 0.42.

However, this is not a realistic situation. Real-life surveys do not deal with populations of size 10. Suppose our population contains 1000 people (still a fairly small population) of which 300 are women and 700 are men.

Now the probability of selecting a woman is 300/1000, which is still 0.3. So if we *can* select the same person twice, the probability that a sample contains two women is $0.3 \times 0.3 = 0.09$, by the multiplication rule as before. For a random sample in which we *cannot* select the same person twice,

$$P(\text{random sample of size 2 contains two women})$$
$$= \frac{300 \times 299}{1000 \times 999} = 0.090 \quad \text{(rounded to three decimal places).}$$

This is essentially the same as for the simple case when we just multiply 0.3 by 0.3. The probability of two men is $(700 \times 699)/(1000 \times 999) = 0.490$ rounded to three decimal places, again the same as the simple case. This is because with a large population, the probability of selecting the same individual twice is so small that it can be ignored.

For the rest of this module we shall assume that populations are large enough for us to apply the multiplication rule, and that the probability of selecting an individual possessing some characteristic remains the same for each selection of a random sample.

In this section, we have introduced the concept of probability and seen how to calculate probabilities for various events, including the composition of random samples of size 2. In the next section, we shall look at probabilities for larger random samples, and relate these to the question on truancy that we posed in Section 1.

Exercises on Section 2

Exercise 2 Colouring probabilities

The hair colour and eye colour of a random sample of 6800 German men are recorded in Table 6.

Table 6 Hair colour and eye colour of 6800 men

		Hair colour				
		Brown	Black	Fair	Red	Total
Eye	Brown	438	288	115	16	857
colour	Grey or green	1387	746	946	53	3132
	Blue	807	189	1768	47	2811
	Total	2632	1223	2829	116	6800

(Source: Goodman, L.A. and Kruskal, W.H. (1954) 'Measures of association for cross classifications', *Journal of the American Statistical Association*, vol. 49, no. 268, pp. 732–64)

Suppose that one of the men in the sample is picked at random. What is the probability that the man:

(a) has blue eyes?

(b) has brown hair?

(c) has blue eyes and brown hair?

(d) has brown hair or black hair?

(e) does not have black hair?

Exercise 3 Saturday afternoon

The probability that Sue will go to watch her favourite hockey team play hockey on Saturday is 0.3. The probability that her team will win is 0.4. What is the probability that Sue will watch her team play on Saturday and they will win?

3 Probability distributions from random samples

At the end of Section 1 we began to investigate the median truancy rate of large schools in the East of England by taking a random sample of the schools. In the sample we found that there were nine values above the (assumed) population median, and three values below. This led us to ask the following question.

> *How likely is it that when selecting a random sample of size 12 from some population, the resulting data will have nine values above the (actual) population median and three below it?*

This is really a question about probabilities, and we can answer it by applying the rules for adding and multiplying probabilities that we studied in Section 2. Although the question refers specifically to nine values above and three values below the population median, our approach will enable us to calculate probabilities for all possible outcomes for any size of random sample. To this end, we need other results for counting the number of ways in which an event can occur.

3.1 Counting combinations

Example 5 Chairperson and secretary

Suppose a club has ten members and we want to know the number of ways in which a chairperson and a secretary for the club could be chosen, if they cannot be the same person. If the chairperson is chosen first, then there are 10 choices of chairperson. There remain nine people from whom to choose the secretary, so the total number of ways of choosing a chairperson and secretary is

$$10 \times 9 = 90.$$

If the club members were the ten people in Example 4 (Ashia, Brenda, Clare, ...), then the 90 choices would be (A, B), (B, A), (A, C) etc.

Note that it makes no difference whether the chairperson or secretary is picked first. If the secretary were picked first, then there would be 10 choices for the secretary, and nine people would remain from whom to pick the chairperson. Again the number of choices for secretary and chairperson would be $10 \times 9 = 90$.

The Programme Committee for the 2012 European Conference on Quality in Official Statistics going to lunch ... or trying to avoid being appointed as chairperson?

Example 6 Chairperson, secretary and treasurer

Suppose the club with ten members want a treasurer as well as a chairperson and secretary. If the chairperson is chosen first, then there are 10 choices of chairperson. If the secretary is chosen next, then there are 9 choices of secretary. There remain eight people from whom to choose the treasurer, so the total number of ways of choosing a chairperson, secretary and treasurer (if they must be different people) is

$$10 \times 9 \times 8 = 720.$$

Activity 21 Chairperson, secretary, treasurer and vice-chairman

Suppose the club with ten members want a chairperson, secretary, treasurer and also a vice-chairperson. If they must all be different people, in how many different ways can they be chosen?

Generalising from these examples, we have the following result.

Number of ways of choosing an ordered sample from a set of objects

Suppose there are n objects to choose from. Then the number of ways of choosing x objects in a specified order is

$$n \times (n-1) \times \cdots \times (n-x+1).$$

(There are x terms in $n \times (n-1) \times \cdots \times (n-x+1)$.)

What we would actually like to count is the number of ways we can select, say, three people from ten *when the order in which the people are selected does not matter*. For example, if a committee of three people is to be formed and they will not have individual roles, then choosing A, then C and then D as the committee would be the same as choosing C, then A and then D.

> *So, how many ways can we form a committee of three people from ten members?*

To tackle this question we can think of the task of choosing a chairperson, secretary and treasurer as a two-stage procedure: first we select a committee of three people, and then we allocate those three people the roles of chairperson/secretary/treasurer. Now the number of ways of allocating three roles to three people is $3 \times 2 \times 1$, as there are three choices of which committee member to make chairperson, then two committee members left for choosing the secretary, and then the one remaining committee member becomes treasurer. Thus, for every choice of committee, there are $3 \times 2 \times 1$ ways of choosing a chairperson, secretary and treasurer. Consequently,

$$\begin{aligned}&\text{number of ways of choosing a chairperson, secretary and treasurer}\\&\quad = 3 \times 2 \times 1 \times (\text{number of ways of choosing a committee of three people}).\end{aligned}$$

Also, from Example 6, we know that the total number of ways of choosing a chairperson, treasurer and secretary from ten people is $10 \times 9 \times 8$. Thus

$$\begin{aligned}&10 \times 9 \times 8\\&\quad = 3 \times 2 \times 1 \times (\text{number of ways of choosing a committee of three people}).\end{aligned}$$

So the number of ways of choosing a committee of three people from ten members is

$$\frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120.$$

Activity 22 Committee of four people

Suppose a club has 12 members.

(a) In how many ways can a chairperson, vice-chairperson, secretary and treasurer be chosen if no person can hold more than one role?

(b) Suppose we have selected four people. In how many ways can they be allocated the roles of chairperson, vice-chairperson, secretary and treasurer?

(c) Hence, what is the number of ways in which a committee of four people can be selected, if the order of selection does not matter?

If we make a selection, and the order in which the people or items are selected does not matter, then the selection is referred to as a **combination**. If there are n people, we let ${}^{n}C_x$ (read as 'n choose x') denote the number of ways of choosing a *C*ombination of x people.

Thus ${}^{10}C_3$ is the number of ways of choosing, from 10 members, a committee of size 3. Similarly, ${}^{15}C_6$ is the number of ways of choosing, from 15 objects, a collection of 6 objects if the order in which they are chosen is of no importance. We have that

$$ {}^{10}C_3 = \frac{10 \times 9 \times 8}{3 \times 2 \times 1}. $$

Note that there are three terms in both the numerator and denominator of ${}^{10}C_3$. Also,

$$ {}^{15}C_6 = \frac{15 \times 14 \times 13 \times 12 \times 11 \times 10}{6 \times 5 \times 4 \times 3 \times 2 \times 1}, $$

where now there are six terms in both the numerator and denominator.

Factorial notation

You may have come across factorial notation in other modules and know that

$$ 3! = 3 \times 2 \times 1, $$

$$ 7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1, \quad \text{and} $$

$$ 10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1. $$

In factorial notation,

$$ {}^{10}C_3 = \frac{10!}{7!\,3!} $$

and it would often be written in this way. However, factorial notation will not be used in this module.

The following summarises the results we have obtained so far in this subsection.

Number of ways of choosing an unordered sample from a set of objects

Suppose there are n objects to choose from. Then the number of ways of choosing x objects if the order does not matter is

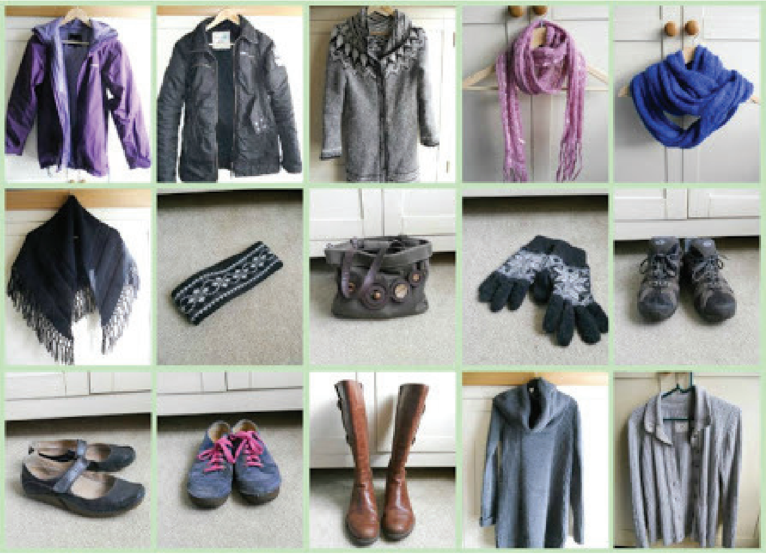
$$ \begin{aligned} {}^{n}C_x &= \frac{\text{number of choices of } x \text{ objects if order does matter}}{\text{number of ways in which } x \text{ objects can be ordered}} \\ &= \frac{n \times (n-1) \times \cdots \times (n-x+1)}{x \times (x-1) \times \cdots \times 1}. \end{aligned} $$

(There are x terms in both $n \times (n-1) \times \cdots \times (n-x+1)$ and $x \times (x-1) \times \cdots \times 1$.)

For any value of n, ${}^nC_0 = 1$ and ${}^nC_n = 1$. This can be seen directly from the definition, because there is only one way that zero objects can be selected from n objects and only one way that n objects can be selected from n objects.

Activity 23 Combinations

Calculate ${}^8C_3$, ${}^7C_5$ and ${}^4C_1$.



More combinations? (of clothes, that is)

3.2 Probabilities of combinations

Here we again suppose that we have a sample of observations from a population. How many of these observations are greater than the population median will partly be a matter of chance. In this section we aim to determine the probability that the number will be 1, the probability that it will be 2, the probability it will be 3, and so forth. This will answer the question posed at the start of Section 3:

> *If we have a sample of size 12, what is the probability that exactly nine values will exceed the population median?*

When the size of the sample exceeds 5 or 6, we will need our results about counting combinations in order to determine the probabilities we want. We shall start, though, by considering the simplest possible case: a random sample of size 1. What is the probability that the selected value lies above the population median? We shall write this probability as $P([+])$, and we shall write $P([-])$ for the probability that the selected value lies below the population median. Using the definition from Section 2, we obtain

$$P([+]) = \frac{\text{number in population above median}}{\text{total number in population}}.$$

The definition of the population median is that half the population lies above it and half below it, so

$$P([+]) = \tfrac{1}{2}.$$

Similarly, $P([-]) = \frac{1}{2}$.

Since the selected value must lie either above or below the median, one of these events must happen; that is, it must be the case that

$$P([+]) + P([-]) = 1.$$

We are ignoring the possibility that the value is exactly equal to the population median.

Our calculated probabilities agree with this $\left(\frac{1}{2} + \frac{1}{2} = 1\right)$.

Moving on to a sample of size 2, let $([+], [-])$ mean that the first observation was $[+]$ and the second was $[-]$. Then

$$P([-],[-]) = P([-]) \times P([-]) = \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4},$$
$$P([-],[+]) = P([-]) \times P([+]) = \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4},$$
$$P([+],[-]) = P([+]) \times P([-]) = \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4},$$
$$P([+],[+]) = P([+]) \times P([+]) = \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4}.$$

If we let $P(x[+])$ denote the probability that the number of $[+]$ is x, then $P(0[+]) = P([-],[-]) = \frac{1}{4}$ and $P(2[+]) = P([+],[+]) = \frac{1}{4}$, while

$$P(1[+]) = P([-],[+]) + P([+],[-]) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}.$$

Exactly one of the events $0[+]$, $1[+]$ and $1[+]$ happens, so

$$P(0[+]) + P(1[+]) + P(2[+]) = \tfrac{1}{4} + \tfrac{1}{2} + \tfrac{1}{4} = 1.$$

A **probability distribution** gives the probability of each possible event that could occur.

When the number of possible outcomes is small, a probability distribution is easily presented as a table or a bar chart. Figure 4 shows the probability distributions that were calculated above for a sample of size 1 and a sample of size 2.

Sample of size 1

Number of $[+]$s, x	0	1
$P(x[+])$	$\frac{1}{2}$	$\frac{1}{2}$

Sample of size 2

Number of $[+]$s, x	0	1	2
$P(x[+])$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

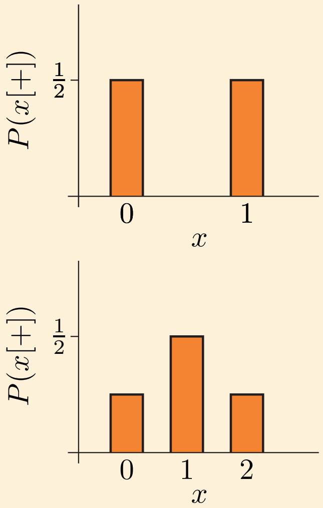


Figure 4 Probability distributions for samples of size 1 and size 2

Activity 24 Probability distribution for a sample of size 3

Suppose a sample of three observations is to be taken from a population.

(a) For each of the four numbers 0, 1, 2, 3, determine the probabilities that the number of observations greater than the median will be the same as it. Check that these four probabilities sum to 1.

(b) Present these values in a table and draw a bar chart of the probability distribution.

Calculations to obtain the probability distribution get longer as the sample size increases. Hence, for samples larger than 3 we need a better method for finding these probabilities. Let us consider the case where the sample size is 5.

Now the probability of the sequence $[+], [-], [-], [+], [+]$ equals

$$\tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^5.$$

Similarly, the probability of the sequence $[-], [-], [-], [+], [-]$ equals

$$\tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^5.$$

Indeed, the probability of any sequence that we specify will equal $\left(\tfrac{1}{2}\right)^5$. Consequently,

$$\begin{aligned} &P(\text{number of [+]s equals } x) \\ &\quad = (\text{number of sequences that have } x \text{ [+]s}) \times \left(\tfrac{1}{2}\right)^5. \end{aligned}$$

Now to get a sequence that has x [+]s, we must just choose x of the five observations to be [+], leaving the remainder of them to be $[-]$. For example, if x is 2 and we label the five observations A, B, C, D and E, then the two $[+]$ values could be chosen as A and C, for example, or A and E, or D and B, and so on. Hence, the number of sequences that have x [+]s is ${}^5C_x$, so

$$P(\text{Number of [+]s equals } x) = {}^5C_x \times \left(\tfrac{1}{2}\right)^5.$$

More generally, we have the following result.

Suppose we take a random sample of size n. Then

$${}^nC_x \times \left(\frac{1}{2}\right)^n$$

is the probability that exactly x of these observations are greater than the population median.

(The formula above is a special case of a probability distribution known as the 'binomial distribution'.)

You have now covered the material related to Screencast 4 for Unit 6 (see the M140 website).

Example 7 Probability distribution for a sample of size 5

A random sample of size 5 is selected.

(a) What is the probability distribution for the number of values that lie above the population median?

(b) What is the probability that exactly four of the selected values lie above the population median?

(c) What is the probability that at least four values lie above the population median?

For the first question, applying the formula gives:

$$P(0[+]) = {}^5C_0 \times \left(\frac{1}{2}\right)^5 = \left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.031\,25,$$

$$P(1[+]) = {}^5C_1 \times \left(\frac{1}{2}\right)^5 = \frac{5}{1} \times \left(\frac{1}{2}\right)^5 = \frac{5}{32} = 0.156\,25,$$

$$P(2[+]) = {}^5C_2 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4}{2 \times 1} \times \left(\frac{1}{2}\right)^5 = \frac{10}{32} = 0.3125,$$

$$P(3[+]) = {}^5C_3 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^5 = \frac{10}{32} = 0.3125,$$

$$P(4[+]) = {}^5C_4 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4 \times 3 \times 2}{4 \times 3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^5 = \frac{5}{32} = 0.156\,25,$$

$$P(5[+]) = {}^5C_5 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.031\,25.$$

These form the probability distribution, and the following table displays it in a clear form.

Table 7 Probability distribution for a random sample of size 5

Number of [+]s, x	0	1	2	3	4	5
$P(x[+])$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$

For the second question, we can now read off the first probability we require:

$$P(4[+]) = \frac{5}{32}.$$

The answer is quite satisfactory in this form. There is no need to write it as a decimal. If you do, it is reasonable to round it to three decimal places and write $P(4[+]) \simeq 0.156$.

The third question asks for the probability that at least four values in the sample lie above the population median. The only possibilities are four values or five values. As these are mutually exclusive events, we use the addition rule:

$$\begin{aligned} P(\text{at least } 4[+]\text{s}) &= P(4[+] \text{ or } 5[+]) \\ &= P(4[+]) + P(5[+]) \\ &= \frac{5}{32} + \frac{1}{32} \qquad \text{(from Table 7)} \\ &= \frac{6}{32} \\ &\simeq 0.188. \end{aligned}$$

Activity 25 Probability distribution for a sample of size 4

A random sample of size 4 is selected.

(a) Calculate the probability distribution for the number of values that lie above the population median. Check that the probabilities are all positive and add to 1.

(b) What is the probability that:

- two of the selected values lie above the population median and two lie below it?

- all the selected values lie on the same side of the population median?

In the question posed at the end of Section 1, we looked at data from a random sample of 12 large schools in the East of England and asked the question:

> *How likely is it that a random sample of 12 would contain nine values above and three values below the assumed population median?*

We repeated the question at the start of Section 3. You are now able to answer it!

Activity 26 Probability that three out of 12 are below the median

Suppose we take a random sample of size 12. What is the probability that there are exactly 9 values above the median?

Later we will need the complete probability distribution for the number of values that exceed the population median in a random sample of size 12. Determining this would be tedious using a calculator, so a computer would always be used. The probability distribution that would be obtained is given in the form of a diagram in Figure 5. The probabilities in the figure are rounded to three decimal places.

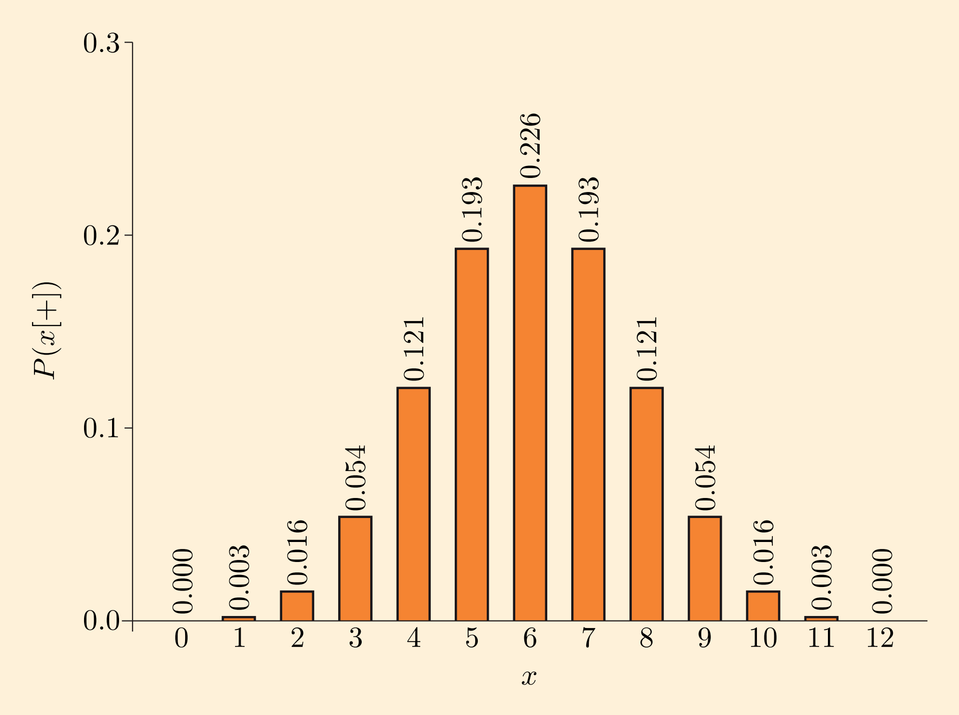


Figure 5 Probability distribution for a random sample of size 12

We can read off from this diagram the probability of any specific outcome. In Section 4 we shall see how to use this information to make an inference about the population from the sample result.

Exercises on Section 3

Exercise 4 Flag signals

A box contains seven flags, each of a different colour. A signal is made by flying three of the flags on a vertical flagpole.

(a) How many different signals can be made? (Note that a signal depends on the order of the flags on the flagpole as well as their colour.)

(b) When a signal is flying, how many different combinations of flags could be left in the box?

Exercise 5 Tossing a coin five times

What is the probability that you get exactly three heads when you toss a coin five times?

4 Testing hypotheses I

In Subsection 1.2 we introduced a measure of truancy: a pupil's truancy rate is the proportion of days that the pupil was absent without authorisation, and a school's truancy rate is the average truancy rate of its pupils. The median truancy rate for all secondary schools in the East of England was 0.98%, and we wondered how the median truancy rate for *large* secondary schools in the East of England compared with this.

In a random sample of 12 large (secondary) schools in the East of England, 9 had a truancy rate above 0.98%. This is more than you would expect if the median for large schools were truly 0.98% – in a sample of 12 you would get 6 values above the median, on average. The question is:

> *Given these sample data, could we reasonably believe that the truancy rate in large schools in the East of England is 0.98%?*

The procedure of using probability to examine this type of question – where sample data are used to evaluate the credibility of a statement – is called a **hypothesis test**. We will first give the steps in a hypothesis test and then discuss them.

Steps in a hypothesis test

1. Make a statement about the population of interest (e.g. *the median truancy rate of large schools in the East of England is 0.98%*). This is the hypothesis we wish to test.
2. Under the assumption that the hypothesis is true, determine the probability distribution for all possible values of some sample statistic (e.g. determine the probability distribution for the number of large schools, out of 12, that will have a truancy rate above 0.98%).
3. Now take the sample and ascertain *how unlikely* the observed value of the sample statistic is, on the basis of (1) and (2) (e.g. are we very unlikely to get a sample statistic as large as 9, that is, a sample of 12 schools in which nine or more schools have a truancy rate above 0.98%?).

4. If the sample turns out to have a very unlikely value, then either:
   - a very unusual event has happened, or
   - the hypothesis suggested in step 1 is incorrect, in which case the sample has provided evidence, albeit in a negative way, that adds something to our beliefs about the population.

Usually the hypothesis stated in step 1 is something that we do not really believe and would like to disprove. This is similar to a trial in a law court, and, indeed, hypothesis tests have much in common with a law-trial:

A person is presumed innocent until proven guilty.

- In a law-trial, the hypothesis is that the person in the dock is innocent. The prosecution do not really believe this, though, which is why the person is in the dock.
- The evidence is examined: a witness identified the defendant as being at the scene of the crime; the defendant has no alibi for the time of the crime; a footprint matches one of the defendant's shoes; etc. The fundamental question is: how likely are these occurrences if the defendant is innocent?
- When the events are extremely unlikely to have arisen *if the defendant is innocent*, it is concluded that the assumption of innocence is wrong, and the defendant is found 'guilty'; i.e. it is concluded that, beyond all reasonable doubt, the hypothesis of innocence is wrong.

There have been cases where probabilities could be calculated under the assumption of a defendant's innocence. This raises the question: in a law trial, when is a probability sufficiently small to equate to 'beyond all reasonable doubt'?

4.1 Tackling the problem

Much of the content in Units 6 to 8 is concerned with making a hypothesis about a population, and then analysing samples to test whether the hypothesis is reasonable. We shall meet several different hypothesis tests; the one we introduce here is called the **sign test** because it is based on the number of $[+]$s and $[-]$s.

Before giving a formal test, the following activity asks you to make your own intuitive judgement about whether sample data indicate that a hypothesis is wrong. The hypothesis is that the median truancy rate of large schools in the East of England is 0.98. The data are the truancy rates in a random sample of 12 such schools.

Activity 27 What is your judgement?

For each of the following instances of random samples of 12 schools, would you think the hypothesis 'The median truancy rate of large schools in the East of England is 0.98' is (i) quite possibly true, (ii) probably wrong, or (iii) almost certainly wrong?

(a) All 12 values are above the hypothesised median.

(b) There is 1 value above and 11 values below the hypothesised median.

(c) There are 6 values above and 6 values below the hypothesised median.

(d) There are 7 values above and 5 values below the hypothesised median.

(e) There are 9 values above and 3 values below the hypothesised median.

The solution to Activity 27 suggests that if all, or nearly all, the values lie on the same side of the assumed median, as in (a) and (b), then we should conclude that the hypothesis is almost certainly wrong. If the values are fairly evenly distributed on both sides of the assumed median, as in (c) and (d), then we should conclude that the hypothesis is quite possibly true. When the outcome is between these two, as in (e), then we are not sure.

Clearly sample data can cast doubt on the truth of a hypothesis. But how much doubt? Just using intuition to judge this is unsatisfactory. Given sample data, we need an objective way of quantifying the evidence that a hypothesis is wrong. In hypothesis testing, we use the probability distribution of all possible outcomes to measure the evidence that a hypothesis is wrong.

Figure 5 gave the probability of each possible outcome when we take a sample of size 12 and the probability of $[+]$ for each item is $\frac{1}{2}$. Figure 5 is reproduced here as Figure 6.

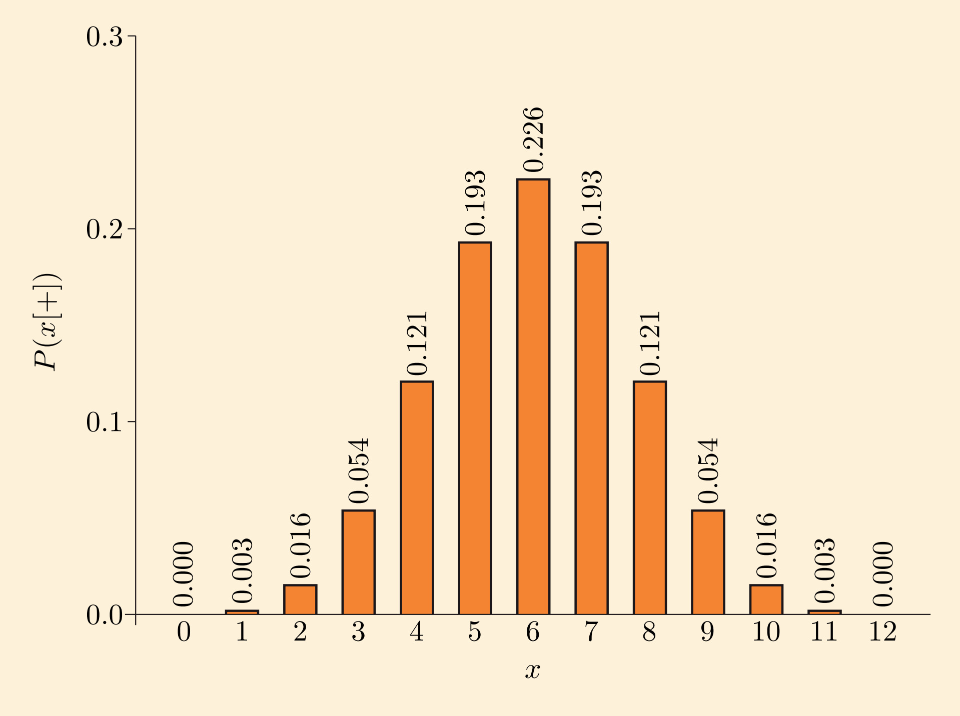


Figure 6 Probability distribution for a random sample of size 12

In a court of law, an unequivocal decision must be made as to whether or not a defendant is guilty. Here, for the moment, we shall assume that we must also decide one way or the other on whether or not to reject the hypothesis about the median. Thus we want a rule – a way of deciding – that tells us how extreme the outcome has to be for us to reject the hypothesis. In Figure 6 you can see that all the individual probabilities are quite small; even the probability of getting 6[+]s and 6[$-$]s is only 0.226 (about $\frac{2}{9}$, so between 1-in-4 and 1-in-5 chance). However, the extreme probabilities are much smaller than this. $P(0[+])$ and $P(12[+])$ are so small they appear as 0.000.

So *if* the hypothesis is true, then the outcome 0[+] is extremely unlikely to occur, as is the outcome 12[+]. This supports our intuitive feeling that if we did select a random sample with either 0[+] or $12[+]$, then we should conclude that the hypothesis is almost certainly wrong. So one possible rule would be to reject the hypothesis if either of these extreme outcomes occurs. However, this rule is erring on the cautious side; it does not cover (b) in Activity 27, which is also very extreme. Another possible rule would be to reject the hypothesis if the sample contains 0[+], 1[+], 11[+] or 12[+].

Activity 28 Probabilistic judgement

(a) Use the probability distribution of Figure 6 to find the probability that a random sample of size 12 has one of the outcomes 0[+], 1[+], 11[+], 12[+].

(b) On the basis of your result in part (a), do you think it is reasonable to reject the hypothesis if a sample is selected with one of these outcomes?

The probability of getting one of these four extreme outcomes is only 0.006 – that is, 6 chances in 1000 or about 1 in 170. Since this probability is also very small, perhaps we are still being over-cautious if we decide to reject the hypothesis only if one of these four outcomes occurs. It might be sensible to incorporate the next most extreme outcomes, 2[+] and 10[+], into our rejection rule, and possibly also 3[+] and 9[+].

When you first come across the idea, you might think it sensible to reject the hypothesis *only* if there is a very small probability of the observed result. For example, you might think it a good idea to reject if all 12 observations are on the same side of the median. However, this means that we would not reject the hypothesis if 11 observations were above the median and only one below it. In such a case, it is much more likely that the hypothesis is wrong and the median is in fact larger than 0.98%. It is a matter of compromise; if we are over-cautious, we will often not reject a hypothesis which is false. We have to remember that we are working with probabilities when we look at samples. *We cannot draw conclusions with certainty; there will always be some doubt.*

How far should we continue? We shall go on adding extreme values into our rejection rule until their combined probability is too large for us to feel justified in rejecting the hypothesis. What is an acceptable value for this probability? This depends on many factors like the importance of the decision, and we shall discuss this in later units. For the present, we shall choose a probability of 0.05 or 5%. This value is used very frequently in many practical situations, and is universally accepted as a reasonable choice.

Significance levels

We refer to 5% as the **significance level** of our hypothesis test.

If a sample is selected whose values are one of the 5% most extreme outcomes that might occur if the hypothesis were true, then *we reject the hypothesis at the 5% significance level.*

Note that this statement does not say ‘we reject the hypothesis with absolute certainty’. Rather, the statement gives the probability that rejecting the hypothesis would be the wrong decision.

Use of the 5% significance level is illustrated in Figure 7; this is a general picture for any size of sample. The actual probabilities will be different for different sample sizes, but the general pattern is the same. The horizontal axis represents the number x of [+]s in the sample, and the vertical axis represents the probability of obtaining different values of x. The 5% most extreme values are shaded.

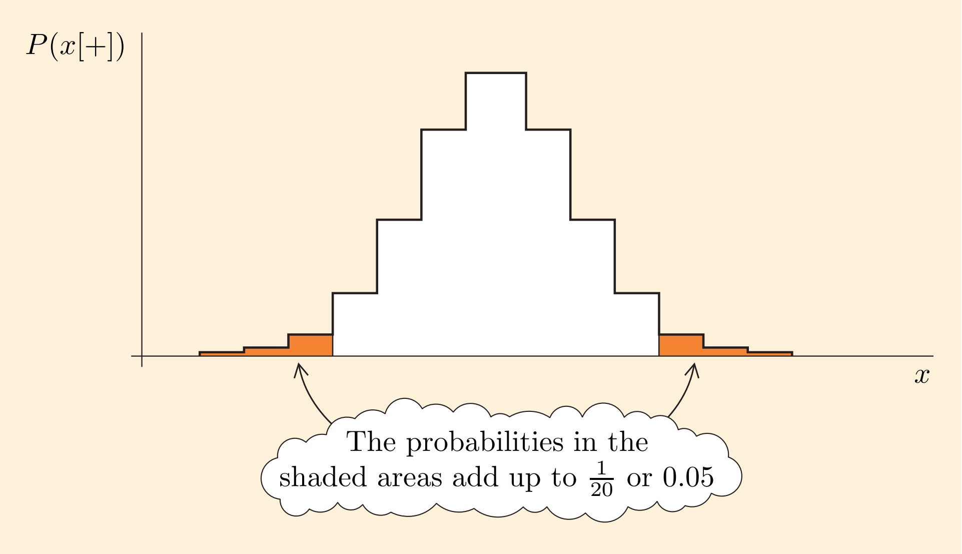


Figure 7 Tail areas adding to 5%

We shall now apply the rule to our random sample of size 12. First, notice that the probability distribution is symmetric, so we can start by considering the left-hand side only, as indicated by the shading in Figure 8. This means finding the outcomes in the left-hand tail whose probabilities add up to $\frac{1}{40} = 0.025$ or $2\frac{1}{2}\%$. Because the distribution is symmetric, there will be a corresponding 0.025 of probability in the right-hand tail, and the two taken together will add up to the required 0.05 or 5%, as shown in Figure 8.

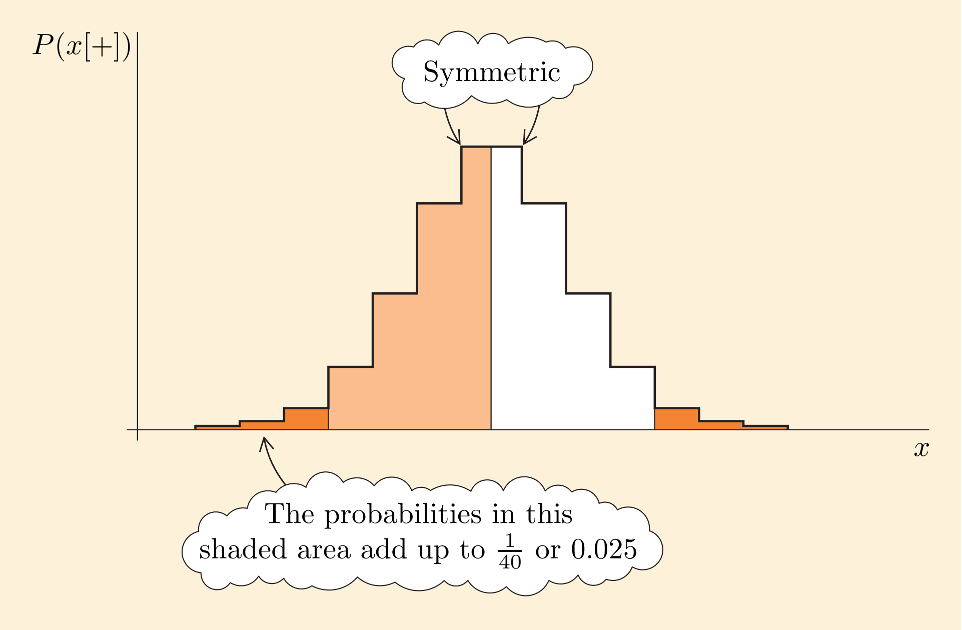


Figure 8 Symmetric distribution

Using the probability distribution of Figure 6 and the addition rule, we find the following probabilities for a random sample of size 12:

$$\begin{aligned}
&P(0[+]) \simeq 0.000,\\
&P(0[+] \text{ or } 1[+]) = P(0[+]) + P(1[+]) \simeq 0.000 + 0.003 = 0.003,\\
&P(0[+] \text{ or } 1[+] \text{ or } 2[+]) \simeq 0.003 + 0.016 = 0.019,\\
&P(0[+] \text{ or } 1[+] \text{ or } 2[+] \text{ or } 3[+]) \simeq 0.019 + 0.054 = 0.073.
\end{aligned}$$

Notice that we only have to add the next probability to the previous total.

The probability of 0, 1 or 2 [+]s is 0.019 which is less than 0.025, whereas if we include 3[+]s the probability is greater than 0.025; neither is exactly equal to 0.025. We shall err on the side of caution and take the value which is less than 0.025. So we shall reject the hypothesis if a sample is selected with one of the outcomes 0[+], 1[+] or 2[+].

We shall also reject the hypothesis at the 5% significance level if the outcome is one of the corresponding extremes on the right-hand side, that is 10[+], 11[+] or 12[+]. To emphasise the symmetry, it is easier to think of these outcomes as 2[$-$], 1[$-$] and 0[$-$]. So for a sample of size 12, we shall reject the hypothesis if the number of values on one side of the assumed median is 2 or fewer. If our hypothesis is true, the probability that there are two or fewer values on one side of the median is $2 \times 0.019 = 0.038$. The probability that there are three or fewer values on one side of the median is $2 \times 0.073 = 0.146$ (about $\frac{1}{7}$). If we decided to reject the hypothesis when we observed three or fewer values on one side of the median, we would have a one in seven chance of rejecting the hypothesis when it was correct. Most people consider this an unacceptably high chance, and so we decide to reject the hypothesis if two or fewer values are on one side of the median.

The number 2 is called the **critical value at the 5% significance level** for a sample of size 12. We can use it to test a hypothesis that any random sample of size 12 comes from a population with some assumed median M. We simply count the number of values above and below M and check whether the smaller number of values is 2 or fewer; if it is, we reject the hypothesis.

This procedure does not apply only to random samples of size 12; for other sample sizes, we proceed in just the same way, though the critical value is different. In the next activity, you will find the critical value for a sample of size 15.

Activity 29 Sample of 15 schools

Suppose that we want to test the hypothesis that the population median is equal to 0.98%, and that we have a random sample of size 15 available. The probability distribution for this situation is shown in Figure 9.

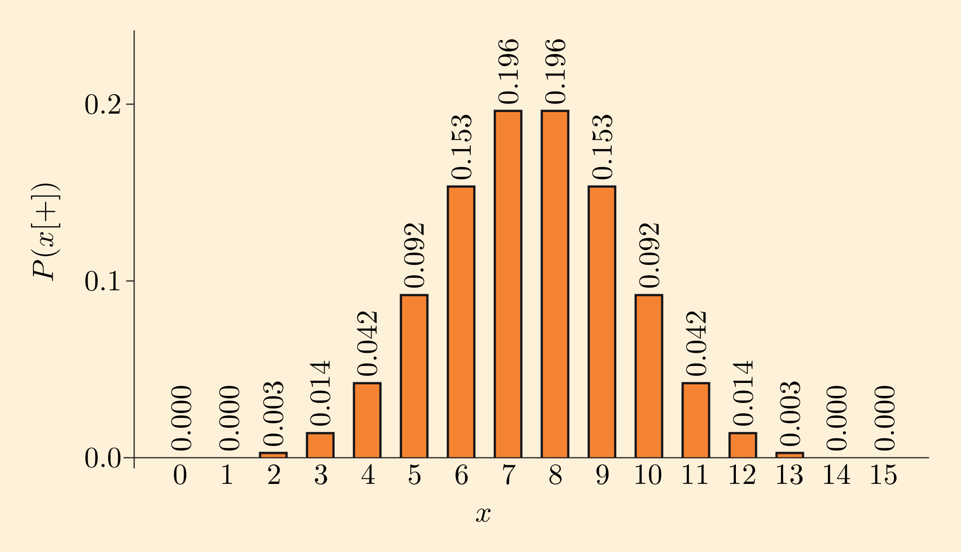


Figure 9 Probability distribution for a random sample of size 15

(a) Calculate the following probabilities:

- $P(0[+]) + P(1[+])$
- $P(0[+]) + P(1[+]) + P(2[+])$
- $P(0[+]) + P(1[+]) + P(2[+]) + P(3[+])$
- $P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) + P(4[+])$

(b) Which of your answers in part (a) is the largest value below 0.025?

(c) Write down the 5% most extreme outcomes, and state the critical value at the 5% significance level.

(d) Would you reject the hypothesis at the 5% level if the sample contained 14 values below the assumed median and 1 above it?

(e) Would you reject the hypothesis at the 5% level if the sample contained 5 values below the assumed median and 10 above it?

(f) Would you reject the hypothesis at the 5% level if the sample contained 12 values below the assumed median and 3 above it?

In Activity 29 you found that the critical value at the 5% significance level for a random sample of size 15 is equal to 3. In the same way, you could calculate the critical value for any size of sample, though you are not expected to do that in this module.

A typical probability distribution is shown in Figure 10. The shaded region corresponds to the outcomes

$$0[+], 1[+], \ldots, C[+]$$

and

$$C[-], \ldots, 1[-], 0[-],$$

where C is the critical value at the 5% significance level. The shaded area is called the **critical region at the 5% significance level**.

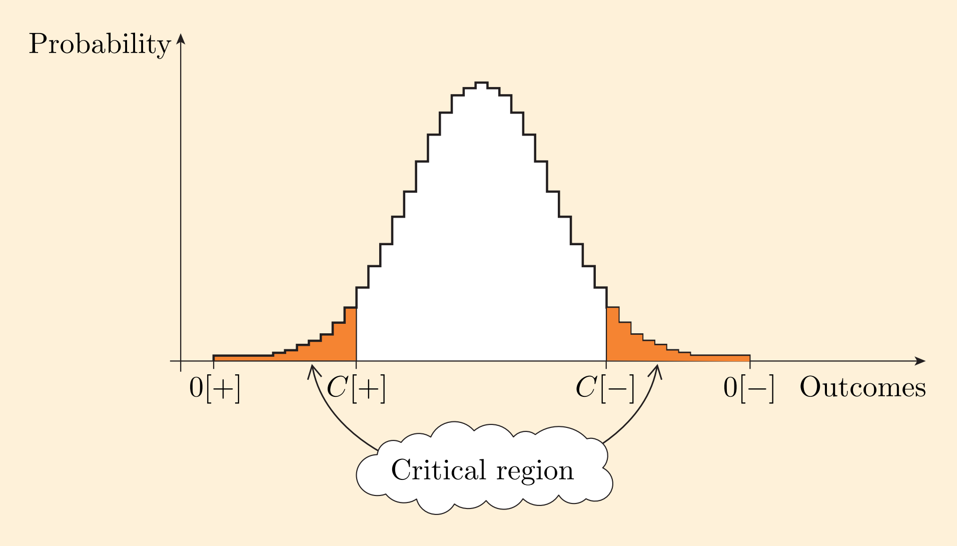


Figure 10 Critical region

Critical region and critical value

The critical *region* at the 5% level is chosen so that the combined probabilities of the outcomes falling in the region is 0.05 or just less than that.

The critical *value* at the 5% significance level can be used to determine whether or not an outcome is in the critical region.

Remember that for a sample of size 15, 2[+] and $13[-]$ describe the same outcome.

For a sample of size 15, the critical value is $C = 3$, and the critical region contains the outcomes $0[+]$, $1[+]$, $2[+]$, $3[+]$, $0[-]$, $1[-]$, $2[-]$ and $3[-]$. From Activity 29, it follows that the combined probability of these outcomes is $2 \times 0.017 = 0.034$, which is less than 0.05.

The procedure can be used to calculate critical values and critical regions for samples of any size. Figure 10 illustrates the critical region for a moderately large sample.

The critical values at the 5% significance level have been calculated for random samples of sizes 1 to 40 and are given in Table 8. (The table could be extended to sample sizes greater than 40, but that is unnecessary for this module.)

Table 8 Critical values at the 5% significance level

Sample size	Critical value at the 5% significance level	Sample size	Critical value at the 5% significance level
1	–	21	5
2	–	22	5
3	–	23	6
4	–	24	6
5	–	25	7
6	0	26	7
7	0	27	7
8	0	28	8
9	1	29	8
10	1	30	9
11	1	31	9
12	2	32	9
13	2	33	10
14	2	34	10
15	3	35	11
16	3	36	11
17	4	37	12
18	4	38	12
19	4	39	12
20	5	40	13

For the sample sizes 1, 2, 3, 4 and 5, no critical value is given. To see why this is the case, consider a sample size of 4, for example. We found the appropriate probability distribution in Activity 25 (Subsection 3.2). The probability of $0[+]$ is $\frac{1}{16} = 0.0625$, the same as the probability of $4[+]$. So the probability of the most extreme outcomes, $0[+]$ or $4[+]$, is $\frac{1}{8} = 0.125$, which is greater than 0.05. So no sample is extreme enough for us to be able to reject the hypothesis at the 5% significance level. The same situation applies for sample sizes 1, 2, 3 and 5.

Activity 30 Critical values

Use Table 8 to find the critical value at the 5% level for the following sample sizes. For each sample, specify the outcomes for which the hypothesis would be rejected at the 5% significance level.

(a) Sample size 21 (b) Sample size 32 (c) Sample size 7

So we can easily decide whether a given sample outcome should lead us to reject a hypothesis at the 5% significance level. We simply count up the number of [+]s and the number of $[-]$s in the sample. If the smaller of these two numbers is less than or equal to the critical value, then we reject the hypothesis.

4.2 The sign test

In Subsection 4.1, we introduced the hypothesis test known as the sign test. You will meet other hypothesis tests later in this module. In this subsection we describe the procedure for the sign test.

All the information we require from the sample is the smaller of the numbers of [+]s and [$-$]s it contains. This number is called the **test statistic**. If the test statistic is less than or equal to the critical value, then we reject the hypothesis at the 5% significance level.

Suppose we want to test whether some population has a median value of M. Then we select a random sample and apply the following procedure.

Procedure: the sign test

1. State the hypothesis that the population median is M.
2. Count the number of values in the sample that are larger than M (denoted by [+]s) and the number of values that are smaller than M (denoted by [$-$]s). The smaller of these two values is the test statistic.
3. Use Table 8 (at the end of Subsection 4.1) to write down the critical value at the 5% significance level corresponding to the size of the sample.
4. Compare the test statistic with the critical value. If it is less than or equal to the critical value, then the hypothesis is rejected at the 5% significance level. If the test statistic is greater than the critical value, then the hypothesis is not rejected.

The symbols $\leq$ and $>$ are often useful in such contexts.

These steps can be summarised by the flow chart given in Figure 11.

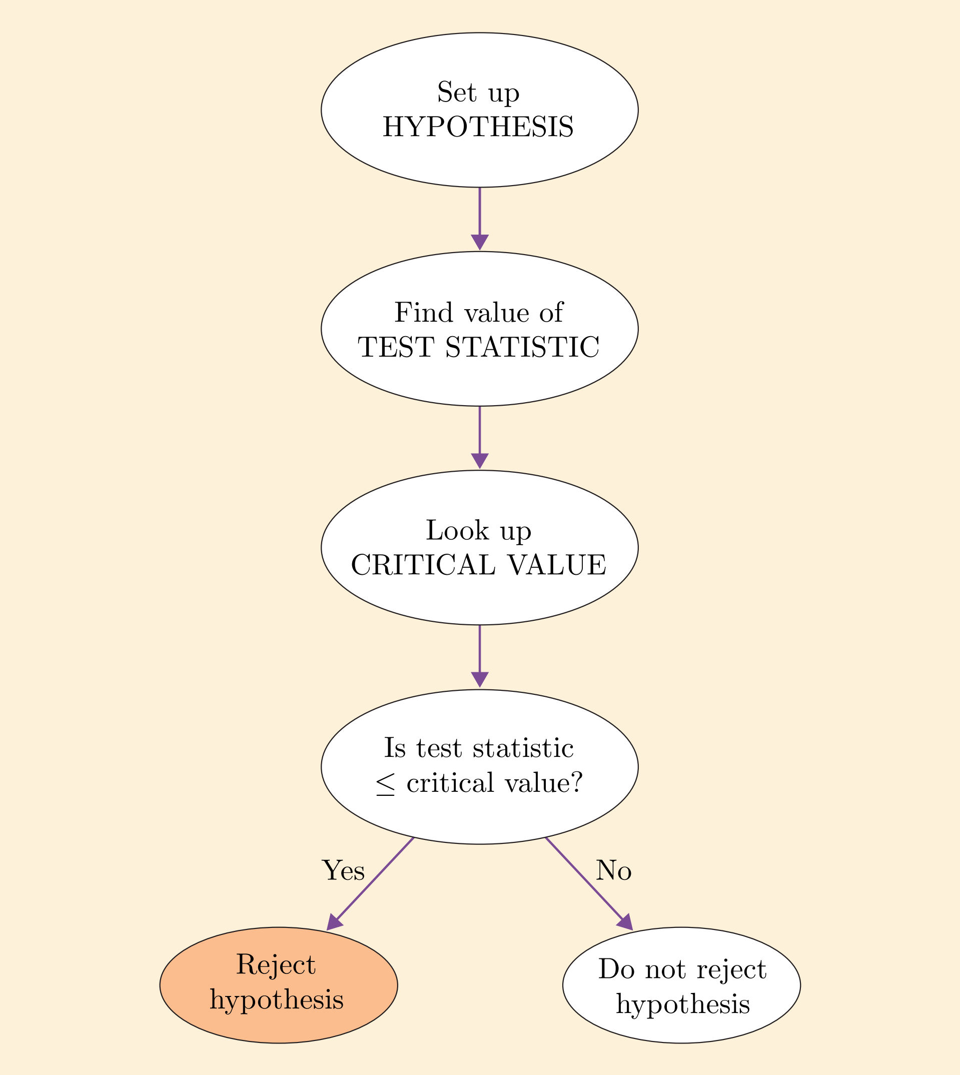


Figure 11 Steps in the sign test

To illustrate how to apply the sign test, let us complete the example we started in Section 1 and continued at the beginning of this section. We wanted to find out whether the truancy rate in large secondary schools in the East of England (those with over 1000 pupils) was the same as the overall rate for all secondary schools in the region. The truancy rate was recorded for a random sample of 12 large schools. We want to test whether this sample comes from a population whose median is 0.98%, the overall truancy rate for the East of England.

So we set up the hypothesis that the population median rate for large schools in the East of England is indeed 0.98%.

In the sample, we count the number of values above and below the assumed median of 0.98%. This gives 9 [+]s and 3 [−]s. The smaller of 9 and 3 is 3, so the test statistic is 3.

From Table 8, the critical value at the 5% significance level for a random sample of size 12 is equal to 2.

We then compare the test statistic with the critical value.

Since 3 is greater than 2, we cannot reject the hypothesis at the 5% significance level. This means that the sample we have observed could quite possibly have been drawn from a population with a median of 0.98%; there is insufficient evidence to suggest otherwise.

So hypothesis testing provides a method of inferring back from the sample to the population. However, it only enables us to reach a statistical conclusion, and there is more to interpreting results than this. We shall discuss this in Subsection 5.3, but for the moment you should concentrate on the statistical technique.

Activity 31 Sample of 23 small schools

In the example, we looked at data from large schools with over 1000 pupils. The data below show the truancy rates from a random sample of 23 small (secondary) schools (those with fewer than 500 pupils) in the East of England. You are asked to investigate whether the median percentage truancy rate for small schools is the same as the median percentage truancy rate for all schools in the East of England.

0.70 0.73 0.16 1.76 0.95 0.80 1.48 0.96 0.64 2.80 0.52 0.96
0.36 0.10 1.21 0.04 0.83 0.64 0.71 0.16 0.71 0.75 0.71

(a) Write down the hypothesis to be tested.

(b) Record the number of values lying above and the number lying below the assumed median. Hence write down the test statistic.

(c) What is the appropriate critical value at the 5% significance level?

(d) Decide whether you would reject the hypothesis at the 5% significance level.

You have now covered the material related to Screencast 5 for Unit 6 (see the M140 website).

So far, we have applied the sign test to questions about truancy rates. It can be used in very many fields of application. One such application is shown next.

Activity 32 Comparing two corn hybrids

An experiment was conducted to compare two different hybrid lines of corn, Hybrid A and Hybrid B, with 28 plots used. Conditions were uniform within a plot, but differed between plots. To treat the hybrids comparably, each plot was divided in two: Hybrid A was grown on one half and Hybrid B on the other. The differences in yields from the two hybrids (Hybrid A − Hybrid B) are given in Table 9.

Examine whether one hybrid is better than the other by testing the hypothesis that the median difference between the two hybrids is 0.

Table 9 Differences in yield between two corn hybrids

1.7	-1.5	-0.6	-5.6	-1.3	-1.9	-10.3	-2.1	-0.6	0.6
-0.7	-0.5	0.1	0.7	0.9	-1.1	-0.5	-1.0	0.2	
-0.9	-0.2	-0.4	0.2	-0.7	-1.4	-0.4	-0.9	-1.5	

(Data source: Dixon, W. J. and Mood, A. M. (1946) ‘The statistical sign test’, *Journal of the American Statistical Association*, vol. 41, pp. 557–566)

Exercises on Section 4

Exercise 6 Testing a claim about petrol consumption

In Unit 1 (Section 3) we introduced values of petrol consumption for a Honda Civic 1.4i. The stemplot of these data is reproduced in Figure 12.

26	1 7 9
27	
28	1 6
29	6 7
30	1
31	7
32	1 6
33	5
34	1
35	0 2 2 3 3 7 8
36	0 2 3 4 5 7
37	5
38	1 7 8
39	1 3
40	5
41	
42	1

$n = 34$ 26 | 1 represents 26.1 miles per gallon

Figure 12 Stemplot of the petrol consumption data

Suppose that a dealer claimed that this type of car would give 37 miles to the gallon. Use the sign test to examine the truth of this claim.

Exercise 7 Do mice like mirrors?

An experiment was conducted to examine whether mice liked to have a mirror in their cage. There were 15 pairs of cages used. The cages in a pair were linked, and one cage in each pair had a mirror in it. A different mouse was placed in each pair, and the time that it spent in each cage was recorded. Three of the 15 mice spent more time in the cage with the mirror, while the other 12 mice spent more time in the cage without the mirror.

Test the hypothesis that the presence/absence of a mirror does not influence where a mouse spends its time.

Is this mouse one of the three?

5 Testing hypotheses II

In Section 4, you were introduced to hypothesis testing and the sign test in particular. In this section, you will learn more about hypothesis testing. We start with an approach that does not reduce the conclusions down to just ‘reject’ or ‘do not reject’.

5.1 Significance probabilities: p-values

From Table 8 (Subsection 4.1), the critical value at the 5% significance level for a sample of size 20 is 5. Suppose, though, that in a random sample of 20 the number of values above the median (the number of $[+]$ values) is only 3. Then, not only do we reject the hypothesised median at the 5% significance level, but we reject it very comfortably, as 3 is almost as close to 0 as it is to 5. Figure 13 gives the probability distribution for the number of values above the median in a random sample of size 20. From the figure, the probability of three or fewer $[+]$ values is only $0.000 + 0.000 + 0.000 + 0.001 = 0.001$. Similarly the probability of three or fewer $[-]$ values is 0.001. Hence, if we used a 0.2% significance level (rather than a 5% significance level), then the critical region would contain the outcomes $0[+]$, $1[+]$, $2[+]$, $3[+]$, $0[-]$, $1[-]$, $2[-]$ and $3[-]$, as the combined probability of these outcomes is $2 \times 0.001 = 0.002$. Stating that we reject a hypothesis at the 0.2% significance level is a much stronger statement than saying we reject it at the 5% significance level. Consequently, we would often like to be more precise when reporting the result of a hypothesis test, rather than simply saying how it compares with the 5% significance level.

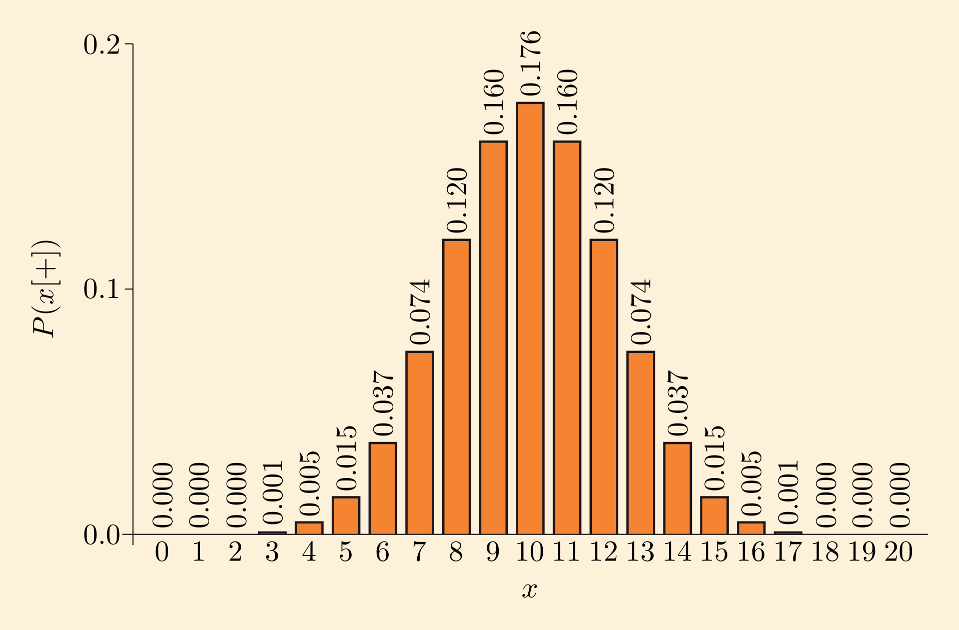


Figure 13 Probability distribution for a random sample of size 20

When we observe three $[+]$ values in a sample of 20, we refer to 0.002 as the **significance probability** or, more commonly, as the ***p*-value** of the hypothesis test. If we set our significance level at anything larger than 0.2% (for example, we might set it to 5%), then we would reject the hypothesis, while if we set our significance level to anything smaller than 0.2%, then we would fail to reject the hypothesis.

Procedure: Obtaining p-values

To obtain the p-value (significance probability) for a hypothesis test, work through the following steps.

1. Assume the hypothesis is true.
2. Consider all the possible outcomes and divide these into two sets:

   Set A contains those outcomes that are as extreme or more extreme than the outcome that actually occurred.

Set B contains those outcomes that are more likely than the outcome that actually occurred.

3. Calculate the probability that a random outcome would be from Set A. This probability is the p-value.

Set A is the smallest critical region that contains the outcome that actually occurred.

When data are analysed on a computer using a standard statistical package, the output almost always reports the result of a hypothesis test as a p-value.

A small p-value indicates that one of the more unlikely outcomes has occurred *or the hypothesis that led to the p-value is false.* As the p-value decreases, such an outcome becomes less likely and the evidence against the hypothesis increases. Table 10 gives a reasonable way of interpreting different p-values. Notice that we never conclude that the hypothesis is true – a large p-value only means that there is little to suggest that the hypothesis is false.

Table 10 Interpretation of p-values

p-value	Rough interpretation
$p > 0.10$	Little evidence against the hypothesis
$0.10 \geq p > 0.05$	Weak evidence against the hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the hypothesis
$0.001 \geq p$	Very strong evidence against the hypothesis

At the end of a hypothesis test it is helpful to give a verbal description of the result and state the associated p-value in order to add precision. For example, we have focused on the hypothesis that the median truancy rate for large schools in the East of England is 0.98%. Using Figure 6, the p-value for the test when there are three $[-]$ values out of 12 is

$$\begin{aligned} &P(0[-]) + P(1[-]) + P(2[-]) + P(3[-]) \\ &\qquad + P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) \\ &\quad = 2 \times [P(0[+]) + P(1[+]) + P(2[+]) + P(3[+])] \\ &\quad = 2 \times (0.000 + 0.003 + 0.016 + 0.054) = 0.146. \end{aligned}$$

We might say, 'The p-value is 0.146, so there is little evidence against the hypothesis that the median truancy rate in large schools in the East of England is 0.98%.' Similarly, if there were only two $[-]$ values out of 12, then the p-value would be $2 \times (0.000 + 0.003 + 0.016) = 0.038$, and we might say, 'The p-value is 0.038, so there is moderate evidence that the median truancy rate in large schools in the East of England is not 0.98%. The data suggest that the truancy rate is higher than that.'

You have now covered the material related to Screencast 6 for Unit 6 (see the M140 website).

Activity 33 Essex schools: testing with p-values

Essex is a county in the East of England region. To compare their truancy rates with the regional median of 0.98%, a random sample of 15 (secondary) schools in Essex was selected and their truancy rates recorded. These rates were as follows:

2.04	1.86	1.64	1.84	0.32	0.62	1.57	1.13
0.95	1.50	0.62	1.46	1.27	0.63	1.44	

(a) What is the hypothesis to be tested?

(b) Record the number of values lying above the median, and the number lying below it. Hence write down the test statistic.

(c) Using Figure 9 (Subsection 4.1), calculate the p-value given by the hypothesis test.

(d) Making reference to this p-value, write down the conclusion to be drawn from the hypothesis test.

Activity 34 Truancy rates in academies

Use Figures 6 and 9 (in Subsection 4.1) for the probabilities you need.

Suppose we took a random sample of secondary school academies in the East of England and examined their truancy rates. For each of the following sample results, test whether the median truancy rate for academies could be 0.98%.

(a) The sample size is 12, with 11 values above 0.98% and 1 value below it.

(b) The sample size is 15, with 3 values above 0.98% and 12 values below it.

5.2 The sign test with ties

Some statisticians deal with ties in other ways. In this module we use the procedure described.

In the examples considered so far, there has never been a case in which one, or more, of the sample values is actually equal to the assumed population median. When this does happen, we shall refer to the situation as a **tie**. This problem is dealt with quite easily: we discard the tied values and reduce the size of the sample accordingly. The procedure is demonstrated in Example 8.

Example 8 Testing with a tie

In this unit, truancy rates have been re-calculated from government figures so as to determine them with an accuracy of two decimal places. In the government publication, the truancy rates are only given to an accuracy of one decimal place. Using the published rates, the median truancy rate for secondary schools in the East of England would be 1.0% (rather than 0.98%).

The data for the 15 schools in Essex (Activity 33) become:

2.0	1.9	1.6	1.8	0.3	0.6	1.6	1.1
1.0	1.5	0.6	1.5	1.3	0.6	1.4	

Suppose we want to use these data to test whether the median truancy rate in Essex differs from a population whose median is 1.0%. There are 10 values above and four values below 1.0%, and one value actually equals 1.0%. We could write this as

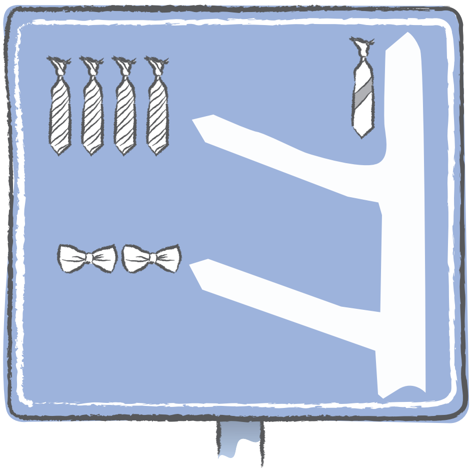
ten [+]s, four [$-$]s and one [=].

We discard the [=] and treat the sample as one of size 14 with ten [+]s and four [$-$]s. So the test statistic is 4. From Table 8 (Subsection 4.1), the critical value is 2. Since 4 is greater than 2, we cannot reject the hypothesis, and we conclude that it is quite possible that the median truancy rate in Essex is 1.0%.

In general, the procedure for dealing with ties is as follows.

Procedure: the sign test with ties

For a sample of size n containing m ties (that is, m of the sample values are equal to the assumed median), discard the m ties and treat the sample as one of size $(n - m)$.



Mark one answer

You are going shopping and need to buy neckties for all your male relatives. Which way should you go?

First left □

Second left □

Straight on □

Do a U-turn □

Another sign test with ties!

Activity 35 Small schools with some tied data

When truancy rates are only recorded to one decimal place, the following are the rates for 23 small schools in the East of England. (The data are from Activity 31.)

0.7	0.7	0.2	1.8	1.0	0.8	1.5	1.0	0.6	2.8	0.5	1.0
0.4	0.1	1.2	0.0	0.8	0.6	0.7	0.2	0.7	0.8	0.7	

We want to test the hypothesis that the truancy rate of small schools in the East of England is 1.0%.

(a) How many values are above the hypothesised median, how many are equal to it, and how many are below it?

(b) What is the test statistic, and what is the sample size that should be used in the hypothesis test?

(c) Test the hypothesis and state your conclusion.

Activity 36 Change in Sheffield temperature?

Meteorological records are available that give the average daily maximum temperature in Sheffield in August for each year from 1883, apart from three years (1918, 1919 and 1923). The median of these temperatures from 1883 to 2011 is $19.7\,^{\circ}$C. The following stemplot gives the average daily maximum August temperatures in Sheffield in the 30 years from 1982 to 2011. If the climate has not changed since 1883, then these temperatures should be a random sample from all the average daily maximum annual temperatures in 1883–2011. That is, the median of these temperatures should be $19.7\,^{\circ}$C.

Will the August temperatures in Sheffield become like those experienced in Florence?

```
17 | 1
17 |
18 | 4 4
18 |
19 | 2
19 | 5 5 6 6 7 7 7 9
20 | 1 1 1
20 | 7 9
21 | 0 1 2 2 3
21 | 5 9
22 | 1 4
22 | 5
23 | 1 3
23 |
24 | 4

n = 30   17 | 1 represents 17.1 °C
```

Figure 14 Stemplot of the average daily maximum August temperature in Sheffield (1982–2011)

We want to test the hypothesis that the values in the stemplot are a random sample from a population whose median is 19.7° C.

(a) How many values are above the hypothesised median, how many are equal to it and how many are below it?

(b) What is the test statistic, and what is the sample size that should be used in the hypothesis test?

(c) Test the hypothesis and state your conclusion.

5.3 Conclusions and reservations

In Sections 4 and 5, the focus has been on truancy rates in different types of (secondary) schools in the East of England, where the rate is 0.98%. We found evidence that small schools in the region had a lower median truancy rate than 0.98%, while there was little evidence that large schools or schools in Essex had a different rate. However, in drawing a conclusion from a hypothesis test, or making any other form of statistical inference, we should think a bit about whether we have any reservations about the analysis. Let us do that now.

First, consider the data we used. Truancy rates for schools were probably obtained from class attendance registers. Presence or absence of each child is noted at the beginning of each morning and afternoon session. However, it is well known that some children attend for the register and then disappear either for the whole session or perhaps just for one particular lesson they dislike.

Also, some schools are probably more disciplined than others in collecting and recording the data. Indeed, because the data are published, it is conceivable that some schools might feel that a low truancy rate would attract new pupils. Other criticisms could also be raised, so we must have some reservations about the accuracy of our data. However, they are from the best publicly available source of information on truancy rates.

Second, let us think about the type of conclusion we drew. We decided that there was no reason to assume that the median truancy rate for large schools in the

East of England was different from 0.98%, which was the overall rate for all schools in the East of England. Notice that we do not say that the rate for large schools is equal to 0.98%. The fact that we do not reject the hypothesis certainly does not justify us assuming that the median is equal to 0.98%. If we examined all the large schools in the East of England, we would almost certainly find that the median truancy rate for large schools differs in the East of England and is not exactly 0.98%. On the basis of the sample we have, though, it is unclear whether the median rate would turn out to be above or below 0.98%.

Note that a statistical test can tell us nothing about causation. For example, we concluded in Activity 31 that the truancy rate for small schools in Essex is probably lower than the overall rate for all schools in Essex. We cannot conclude from this that the small size of a school *causes* the truancy rate to fall. There may well be other factors, like the situation of the school, which influence both the size and the truancy rate.

The reservations expressed in this subsection may have left you feeling that it was not worth doing the statistical analyses. However, the analyses have added to our knowledge about truancy rates. The point is that a statistical analysis enables you to quantify the uncertainty that exists in drawing conclusions – in most other circumstances, the uncertainty of conclusions is either not recognised or is ignored.

'Thanks Mary, that does put our data in perspective.'

Exercises on Section 5

Exercise 8 A second look at mice and mirrors

In Exercise 7 (Section 4) an experiment was described in which 12 out of 15 mice preferred a cage without a mirror, while three mice preferred a cage with a mirror. The exercise asked for a test of the hypothesis that the presence/absence of a mirror does not influence where a mouse spends its time. Using the probabilities given in Figure 9 (Subsection 4.1), determine the p-value of the test and evaluate the evidence against the hypothesis.

Exercise 9 Change in depression

A psychologist rated 14 subjects undergoing withdrawal from narcotics on the basis of the extent of their depression before and one hour after receiving a dose of methadone. The degree of depression was rated on a scale from 1 (= no depression) to 5 (= severe depression). The results are given in Table 11.

Table 11 Depression scores before and after a dose of methadone, and their differences

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Before	2	1	2	3	2	2	2	2	2	1	4	3	2	3	3	2
After	1	1	3	1	1	2	4	1	1	1	2	2	1	1	3	3
Change	-1	0	1	-2	-1	0	2	-1	-1	0	-2	-1	-1	-2	0	1

If methadone does not affect depression, then the difference between a subject's 'before' and 'after' scores is equally like to be above or below 0. (It could also equal 0.) Thus, we want to test the hypothesis that the median difference between 'before' and 'after' scores is 0 – rejecting this hypothesis would show evidence that methadone affects depression.

(a) How many values are above the hypothesised median, how many are equal to it, and how many are below it?

(b) What is the test statistic, and what is the sample size that should be used in the hypothesis test?

(c) Test the hypothesis and state your conclusion.

6 Computer work: probabilities and the sign test

In this section you will use the statistical package Minitab to calculate probabilities of the type that you calculated (for small sample sizes) in Subsection 3.2. You will also learn how to use Minitab to perform a sign test for a population median. You should now turn to the Computer Book and work through Chapter 6.

Summary

In this unit, you have been introduced to the basic concepts of probability and hypothesis testing. These are incredibly important in statistics: probability underpins most statistical methods, and one of the most common tasks in statistical inference is hypothesis testing. A consequence is that the ideas introduced in this unit will be reinforced in subsequent units and, indeed, in any further statistics modules that you study.

You have learned the fundamentals of probability – its definition as a proportion, the addition rule and the multiplication rule. You have also learned how to count the number of ways that combinations can occur, so that you can now calculate the probability distribution for the number of sample data that will exceed the population median. You used this probability distribution to perform a sign test – while simultaneously learning the structure of a hypothesis test, the meaning of a p-value, and how to report your conclusions and consider reservations you might have about the test. The range of tasks you can perform with Minitab has been extended.

Learning outcomes

After working through this unit, you should be able to:

- appreciate how the modelling diagram from earlier units can be modified to take account of the important step of inferring back from the sample to the population
- appreciate that a general question needs to be clarified and made more precise before it can be answered using statistics
- calculate probabilities based on random selection
- apply the addition rule for mutually exclusive events
- apply the probability rule for complementary events
- apply the multiplication rule for statistically independent events
- count the number of ways that a specified combination can occur
- calculate the probability distribution for the number of values above the median of the population
- understand the concepts of a hypothesis test and of rejecting a hypothesis at the 5% significance level
- look up the critical value for the sign test at the 5% significance level, and use this value to determine the critical region
- apply the sign test, including the case when ties are present
- consider what reservations you have about the conclusion drawn from a hypothesis test
- use Minitab to calculate the probability distribution for the number of outcomes that exceed the median and perform a sign test.

Solutions to activities

Solution to Activity 1

There are many possible factors you could have thought of. Some are given here, but your list may be quite different and equally valid.

Work: children may absent themselves from a particular lesson because they dislike that subject, have not completed homework, etc.

Home background: parents who take their children to school and are interested in their education will encourage regular attendance.

Bullying: children may stay away because they are being bullied and are frightened to go to school.

Helping at home: children may be encouraged by their parents to miss school to help with housework or take care of young siblings, or they may feel they need to look after their parents.

Age: truancy is much commoner among older children approaching school-leaving age than among those in primary schools.

Solution to Activity 2

As with the previous activity, there are many possible answers here. Some are listed below.

Level: primary or secondary school

Location: inner city or rural school, for example

Size of school

Type of school: independent, academy, etc.

Solution to Activity 3

There are many possible reasons that are definitely truancy: missing school to go to the cinema/football match/fishing; missing school to avoid boredom/a lesson you hate; etc.

Reasons that would not normally be classed as truancy include illness, medical/dental appointments, attendance at family funerals, being kidnapped, religious observance.

Reasons that might (or might not) be truancy include staying off school because you didn’t feel that well, when it would depend on how unwell you felt and how often it happened. Another example might be staying off school to look after a sick brother/sister/parent or go on holiday with parents, when it might again depend on how often this happened, and might also depend on whether the school had given permission.

Solution to Activity 4

In order to compare truancy in different schools, we must somehow take account of the number of pupils in each school. If we simply counted the number of cases of truancy, then large schools would almost inevitably have higher truancy counts. So we need some measure of the rate of truancy in a school. Here are three possible ways of collecting data and measuring truancy; you may well have thought of others.

(a) Use the school attendance registers which record the children present at the beginning of each morning and afternoon. Record which of the children who were not present had notes of authorised absence due to illness or other reason and discount these cases. Calculate the proportion of children absent without authorisation, averaged over a year.

(b) Choose a particular week. Count the number of children present for each lesson during that week, and divide this figure by the number of lessons and the number of children in the school. Subtract this number from 1 to give a truancy rate.

(c) At some arbitrary time, send someone round the school to count the number of children present. Subtract this number from the total number on the school register to get the number who are absent. Divide by the total number of children to give a truancy rate.

None of these ways is ideal, but you can see that it is not a simple matter to obtain a reliable figure; there are many factors to consider.

Solution to Activity 5

(a) It would be possible to find the median of the 12 values by drawing a stemplot, but with so few, it is probably easier just to list them in order:

0.31, 0.82, 0.83, 1.06, 1.09, 1.19, 1.44, 1.52, 1.84, 1.88, 2.78, 2.90.

Since the sample size is even,

See Subsection 4.2 of Unit 1.

$$\begin{aligned}\text{median} &= \frac{\text{sum of two middle data values}}{2}\\ &= \frac{1.19 + 1.44}{2}\\ &= 1.315 \simeq 1.32.\end{aligned}$$

So the median of the sample is 1.32%.

(b) The median value of the truancy rate of the sample of 12 large East of England schools is somewhat larger than the median truancy rate for all the East of England schools. However, a different sample would probably have had a different median. In summary, the median truancy rate of all large schools in the East of England looks as if it might be larger than 0.98%, but we cannot be sure.

Solution to Activity 6

This is a suggested ranking of the events, from the most likely to the least likely.

C. The sun will rise tomorrow.

D. The sun will shine tomorrow.

F. You toss a coin and it lands tails up.

G. You throw a die and it shows a six.

B. Two out of a group of ten people have the same birthday.

A. Your new colleague at work has the same birthday as you.

E. You will win the jackpot in the National Lottery next week.

H. A member of a hockey team is 150 years old.

Your ranking may be slightly different from this, but it should be fairly similar.

(Note that event C can be regarded as certain to happen, whereas event H is impossible.)

Solution to Activity 7

(a) The probability of selecting a male student is

$$\frac{\text{number of male students}}{\text{total number of students}} = \frac{3293}{6082} \simeq 0.541.$$

(b) The probability of selecting a law student is

$$\frac{\text{number of law students}}{\text{total number of students}} = \frac{334}{6082} \simeq 0.055.$$

(c) The probability of selecting a female medical student is

$$\frac{\text{number of female medical students}}{\text{total number of students}} = \frac{206}{6082} \simeq 0.034.$$

Solution to Activity 8

(a) The probability of selecting a woman is

$$\frac{\text{number of women in population}}{\text{total number in population}}.$$

If we apply this definition to the case where there are 5 men and 5 women, then the probability of selecting a woman is

$$\frac{5}{10} = 0.5.$$

(b) For a population consisting of 1 man and 9 women, using the definition in the solution to (a) gives the probability of selecting a woman as

$$\frac{9}{10} = 0.9.$$

(c) For 9 men and 1 woman, the probability of selecting a woman is

$$\frac{1}{10} = 0.1.$$

(d) For 10 men and 0 women, the probability of selecting a woman is

$$\frac{0}{10} = 0.$$

(e) For 99 men and 1 woman, the probability of selecting a woman is

$$\frac{1}{100} = 0.01.$$

(f) For 0 men and 100 women, the probability of selecting a woman is

$$\frac{100}{100} = 1.$$

Solution to Activity 9

(a) For each of the events, the probabilities are as follows:

$$\begin{aligned}
&P(\text{child absent through truancy for 5 to 9 days})\\
&\quad= \frac{\text{total number absent for 5 to 9 days}}{\text{total number of children}}\\
&\quad= \frac{90}{300} = 0.3.\\
&P(\text{child absent for 5 to 9 days and attends School B})\\
&\quad= \frac{30}{300} = 0.1.\\
&P(\text{child from School A is absent for 10 to 19 days})\\
&\quad= \frac{\text{number absent for 10 to 19 days from School A}}{\text{total number at School A}}\\
&\quad= \frac{26}{200} = 0.13.\\
&P(\text{child absent for} \geq \text{10 days})\\
&\quad= \frac{\text{number absent for 10 to 19 days} + \text{number absent for} \geq \text{20 days}}{\text{total number of children}}\\
&\quad= \frac{45+15}{300} = \frac{60}{300} = 0.2.
\end{aligned}$$

(b) Let C be the event that a child is absent through truancy for 5 to 9 days and let B be the event that a child attends School B. (You might well have chosen different letters: that does not matter.)

Then

$$P(\text{child absent through truancy for 5 to 9 days}) = P(C)$$

and

$$P(\text{child absent for 5 to 9 days and attends School B}) = P(C \text{ and } B).$$

Solution to Activity 10

(a) The events are mutually exclusive, as a person can only have one blood type.

(b) The events are not mutually exclusive. Even though it is not very common, a person can have black hair and blue eyes.

(c) These are not mutually exclusive events. Although all three events cannot simultaneously happen, two of them could happen at the same time. (In fact, any two of them could happen at the same time.)

(d) These are not mutually exclusive events. While (i) and (ii) cannot occur simultaneously, either of them could occur with (iii).

Solution to Activity 11

(a) Altogether there are 200 children at School A. Using the data in Table 3, the probabilities can be calculated as follows.

There are 108 children absent through truancy for 0 to 4 days. Hence

$$P(\text{child from School A absent for 0 to 4 days}) = \frac{108}{200} = 0.54.$$

There are 60 children absent through truancy for 5 to 9 days. Hence

$$P(\text{child from School A absent for 5 to 9 days}) = \frac{60}{200} = 0.3.$$

Altogether $108 + 60 = 168$ children are absent through truancy for between 0 and 9 days. Hence

$$P(\text{child from School A absent for 0 to 9 days}) = \frac{168}{200} = 0.84.$$

(b) Let E denote the event that a child at School A is absent through truancy for 0 to 4 days, and F denote the event that the child is absent for 5 to 9 days. The events E and F are mutually exclusive, as both cannot apply to any one child.

From the solution to part (a), $P(E) = 0.54$ and $P(F) = 0.3$. The addition rule for mutually exclusive events states that

$$P(E \text{ or } F) = P(E) + P(F).$$

Now $P(E) + P(F) = 0.54 + 0.3 = 0.84$, which is the value we found in part (a) for the probability that a child at School A is absent through truancy for 0 to 9 days, that is $P(E \text{ or } F)$, so the addition rule holds in this case.

Solution to Activity 12

(a) Let M stand for the event that a man is selected and W stand for the event that a woman is selected.

$$P(M) = \frac{3293}{6082} \simeq 0.541.$$

Hence, $P(W) = 1 - P(M) \simeq 1 - 0.541 = 0.459$.

(b) $P(\text{selecting a medical student}) \simeq \dfrac{575}{6082} = 0.095.$

Hence

$$P(\text{selected student is not a medical student}) \simeq 1 - 0.095 = 0.905.$$

This is quicker than the alternative of putting:

$$\begin{aligned} &P(\text{selected student is not a medical student}) \\ &\quad = \frac{\text{number of students in science, arts and law}}{\text{number of students}} \\ &\quad = \frac{2602 + 2571 + 334}{6082} = \frac{5507}{6082} \simeq 0.905. \end{aligned}$$

Solution to Activity 13

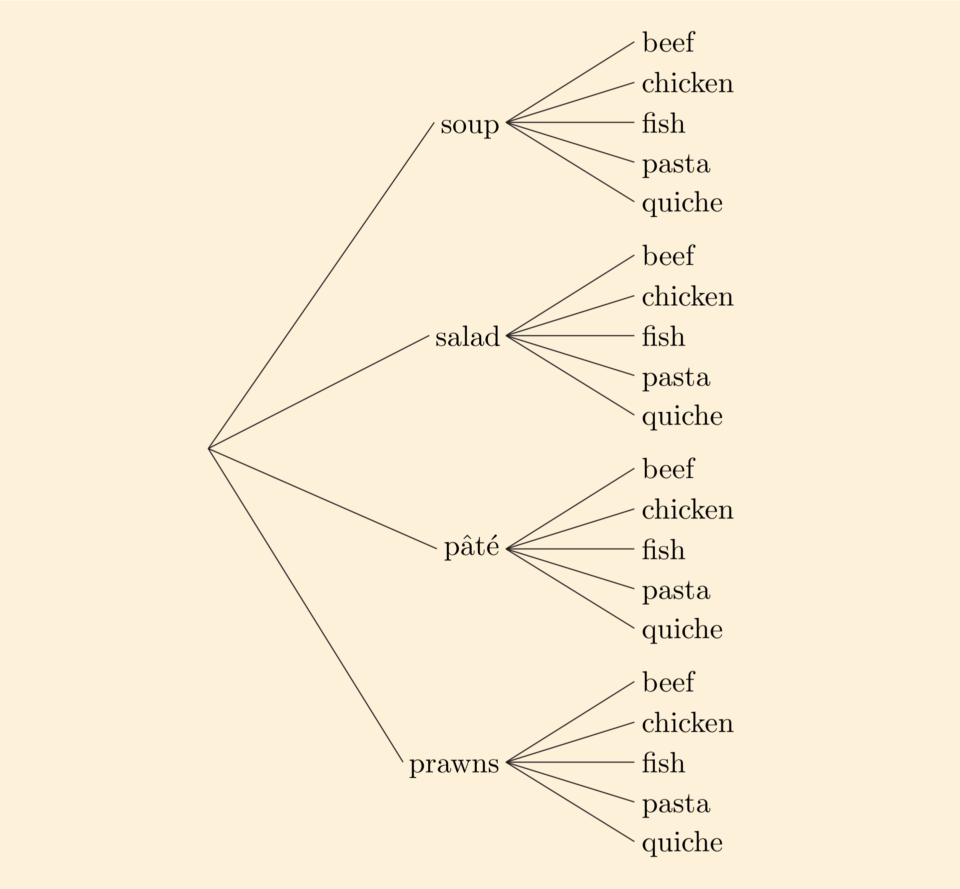
(a) 'G does not occur' is the event that a child selected at random from the two schools is absent through truancy for 0 to 4 days or is absent for 10 or more days.

(b)

$$\begin{aligned} P(G \text{ does not occur }) &= 1 - P(G) \\ &= 1 - \frac{90}{300} = 0.7. \end{aligned}$$

Solution to Activity 14

(a)



Tree diagram for a two-course meal

(b) There are four vegetarian combinations: soup–pasta, soup–quiche, salad–pasta and salad–quiche.

(c)

$$\begin{aligned}&P(\text{vegetarian meal})\\&\quad=\frac{\text{number of vegetarian meal combinations}}{\text{total number of meal combinations}}=\frac{4}{20}=0.2.\end{aligned}$$

(d)

$$\begin{aligned}&P(\text{vegetarian first course})\\&\quad=\frac{\text{number of vegetarian first courses}}{\text{total number of first courses}}=\frac{2}{4}=0.5.\end{aligned}$$

$$\begin{aligned}&P(\text{vegetarian second course})\\&\quad=\frac{\text{number of vegetarian second courses}}{\text{total number of second courses}}=\frac{2}{5}=0.4.\end{aligned}$$

The product of these answers is $0.5 \times 0.4 = 0.2$, which is the answer in (c).

Solution to Activity 15

(a) In part (a) of Activity 14, you found that there are 20 different two-course meal combinations. From each of these two-course combinations we can form three different three-course combinations, because there are three choices for the third course. Hence in total there are $20 \times 3 = 60$ different three-course combinations.

(b)

$$\begin{aligned} &P(\text{third course is good with custard}) \\ &\quad = \frac{\text{number of third courses good with custard}}{\text{total number of third courses}} = \frac{2}{3} \simeq 0.667. \end{aligned}$$

Two of the first courses are vegetarian, as are two of the second courses, and two of the third courses are good with custard. Hence the number of three-course meals that have two vegetarian courses followed by a course that is good with custard is $2 \times 2 \times 2 = 8$. (They are: soup–pasta–pie, soup–quiche–pie, salad–pasta–pie, salad–quiche–pie, soup–pasta–crumble, soup–quiche–crumble, salad–pasta–crumble and salad–quiche–crumble.) Hence

$$\begin{aligned} &P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course} \\ &\quad \textit{and} \text{ third course is good with custard}) \\ &\quad = \frac{8}{60} \simeq 0.133. \end{aligned}$$

(c)

$$\begin{aligned} &P(\text{vegetarian first course}) \times P(\text{vegetarian second course}) \\ &\qquad \times P(\text{third course is good with custard}) \\ &\quad \simeq 0.5 \times 0.4 \times 0.667 \simeq 0.133. \end{aligned}$$

This is the same result as found in part (b) for

$P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course})$.

Solution to Activity 16

(a) The events are not independent. Taller people tend to be heavier, so if a person is taller than average, then they are more likely to be heavier than average.

(b) These events are independent. The height of the first person has no influence on the weight of the second person, assuming the people were chosen at random. (If you choose two people standing next to each other, then the choices are not random, and the characteristics of one person could relate to the characteristics of the other.)

(c) These events are independent, as the day on which you are born has no influence on your weight.

(d) The events are not independent. For example, if the first two events both occur then we *know* that the third event occurs.

(e) The probability that the card is an ace is

$$\frac{\text{number of aces in pack}}{\text{number of cards in pack}} = \frac{4}{52} = \frac{1}{13}.$$

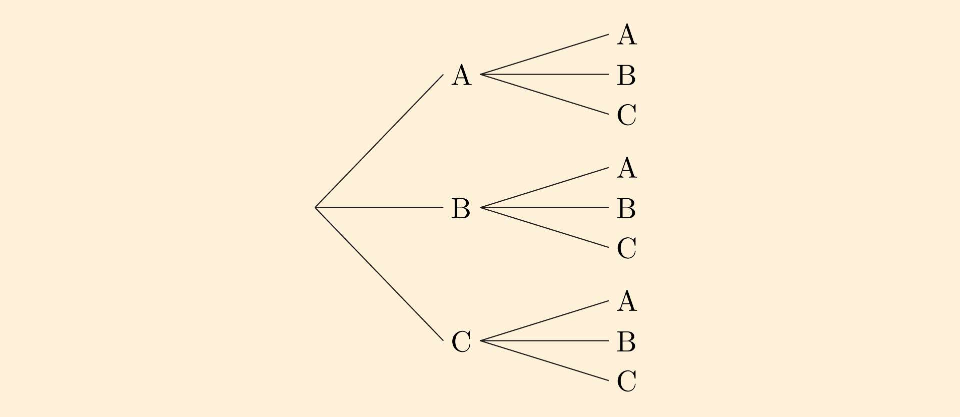
If we know that the card is a diamond, then the probability that it is an ace is

$$\frac{\text{number of cards that are the ace of diamonds}}{\text{number of diamonds}} = \frac{1}{13}.$$

These probabilities are the same. That is, knowing that one of the events occurs does not change the probability that the other event occurs. Hence the events are independent.

Solution to Activity 17

(a) Using A, B and C to represent Ashia, Brenda and Clare, the different possible samples are shown in the following tree diagram. So, the possible samples of size 2 are (A, A), (A, B), (A, C), (B, A), (B, B), (B, C), (C, A), (C, B) and (C, C). Altogether there are 9 possible samples.



(b) Because there are 3 possibilities for the first member of the sample and 3 possibilities for the second member, there are $3 \times 3 = 9$ possible samples altogether.

Solution to Activity 18

We have that $P(M) = 7/10 = 0.7$, so

$$P(2M) = P(M) \times P(M) = 0.7 \times 0.7 = 0.49.$$

Solution to Activity 19

At any selection we can write down the probabilities of selecting a man or a woman:

$$P(M) = \frac{40}{100} = 0.4 \quad \text{and} \quad P(W) = \frac{60}{100} = 0.6.$$

For a sample of size 2,

$$\begin{aligned} P(2M) =& 0.4 \times 0.4 = 0.16, \\ P(0M) =& P(2W) = 0.6 \times 0.6 = 0.36, \\ P(1M) =& P(\text{man selected first and woman selected second}) \\ & \quad + P(\text{woman selected first and man selected second}) \\ =& 0.4 \times 0.6 + 0.6 \times 0.4 \\ =& 0.24 + 0.24 = 0.48. \end{aligned}$$

As a check,

$$P(2M) + P(0M) + P(1M) = 0.16 + 0.36 + 0.48 = 1.00.$$

Solution to Activity 20

We are now considering only samples where the same person cannot be considered twice. There are three women, so the first one can be selected in 3 ways. Then only two women remain, so the second can be selected in 2 ways. Hence the total number of samples of size 2 consisting of two women is $3 \times 2 = 6$.

As a check, the samples of two women are (A, B), (A, C), (B, A), (B, C), (C, A), (C, B), where, for example, (A, B) means that Ashia is selected first and Brenda is selected second.

Solution to Activity 21

If they are picked in the order chairperson, secretary, treasurer, vice-chairperson, then there are 10 choices for chairperson, 9 choices for secretary, 8 choices for treasurer, and 7 choices for vice-chairperson, giving a total of

$$10 \times 9 \times 8 \times 7 = 5040 \text{ choices.}$$

Solution to Activity 22

(a) Here the order of selection matters. The chairman can be chosen from 12 members, the vice-chairman from the remaining 11, etc. The number of ways of choosing people for the 4 roles equals $12 \times 11 \times 10 \times 9 = 11\,880$.

(b) The number of ways of allocating 4 people the different roles is $4 \times 3 \times 2 \times 1 = 24$.

(c) The number of ways of choosing a committee of 4 people from 12 members is

$$\frac{\text{number of choices of 4 people if order of choice matters}}{\text{number of ways of allocating 4 people to 4 roles}} = \frac{11\,880}{24} = 495.$$

Solution to Activity 23

$$\begin{aligned}
{}^8C_3 &= \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = \frac{336}{6} = 56. \\
{}^7C_5 &= \frac{7 \times 6 \times 5 \times 4 \times 3}{5 \times 4 \times 3 \times 2 \times 1} = \frac{2520}{120} = 21. \\
{}^4C_1 &= \frac{4}{1} = 4.
\end{aligned}$$

Solution to Activity 24

(a) The four probabilities required are:

$$\begin{aligned}
P(0[+]) &= P([-],[-],[-]) = P([-]) \times P([-]) \times P([-]) \\
&= \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^3 = \tfrac{1}{8}, \\
P(1[+]) &= P([-],[-],[+]) + P([-],[+],[-]) + P([+],[-],[-]) \\
&= \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^3 = 3 \times \left(\tfrac{1}{2}\right)^3 = \tfrac{3}{8}, \\
P(2[+]) &= P([-],[+],[+]) + P([+],[-],[+]) + P([+],[+],[-]) \\
&= \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^3 = 3 \times \left(\tfrac{1}{2}\right)^3 = \tfrac{3}{8},
\end{aligned}$$

and

$$P(3[+]) = P([+],[+],[+]) = \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^3 = \tfrac{1}{8}.$$

Adding these probabilities together gives

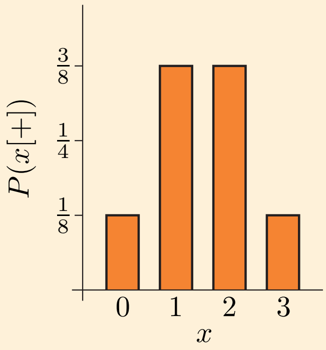
$$P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) = \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{8} = 1,$$

as required.

(b)

Sample of size 3

Number of [+]s, x	0	1	2	3
$P(x[+])$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$



Probability distribution for a random sample of size 3

Solution to Activity 25

(a) Now $n = 4$.

$$P(0[+]) = {}^4C_0 \times \left(\frac{1}{2}\right)^4 = 1 \times \left(\frac{1}{2}\right)^4 = \frac{1}{16}.$$

$$P(1[+]) = {}^4C_1 \times \left(\frac{1}{2}\right)^4 = \frac{4}{1} \times \left(\frac{1}{2}\right)^4 = \frac{4}{16} = \frac{1}{4}.$$

$$P(2[+]) = {}^4C_2 \times \left(\frac{1}{2}\right)^4 = \frac{4 \times 3}{2 \times 1} \times \left(\frac{1}{2}\right)^4 = \frac{6}{16} = \frac{3}{8}.$$

$$P(3[+]) = {}^4C_3 \times \left(\frac{1}{2}\right)^4 = \frac{4 \times 3 \times 2}{3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^4 = \frac{4}{16} = \frac{1}{4}.$$

$$P(4[+]) = {}^4C_4 \times \left(\frac{1}{2}\right)^4 = \frac{4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^4 = \frac{1}{16}.$$

Probability distribution for a random sample of size 4

Number of [+]s, x	0	1	2	3	4
$P(x[+])$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

The probabilities are positive and

$$\frac{1}{16} + \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{1+4+6+4+1}{16} = 1.$$

(b) From the table, the probability that two of the selected values lie above the median (and hence two lie below it) is $\frac{3}{8}$.

If all selected values lie on the same side of the population median, they must be either all above or all below it. Applying the addition rule for mutually exclusive events, we obtain

$$\begin{aligned} P(4[+] \text{ or } 0[+]) &= P(4[+]) + P(0[+]) \\ &= \frac{1}{16} + \frac{1}{16} = \frac{1}{8} \quad (= 0.125). \end{aligned}$$

Solution to Activity 26

$$\begin{aligned}P(\text{number of [+]s equals 9}) &= {}^{12}C_9 \times \left(\tfrac{1}{2}\right)^{12}\\ &\simeq \frac{12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4}{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \times 0.000\,244\\ &= \frac{12 \times 11 \times 10}{3 \times 2 \times 1} \times 0.000\,244 \simeq 0.054.\end{aligned}$$

Solution to Activity 27

There is no definitive answer to this question. Your answer may well differ from that of the module team, but it should follow the same pattern.

In the population of all schools, by definition, half will have a truancy rate above the median and half will have one below it. As a random sample is representative of the population, we might expect this property to apply in the sample to some extent.

(a) If all 12 values were above the assumed median, we should be very surprised. It is much more likely that the hypothesised median is not the real median. The module team would conclude that the hypothesis is almost certainly wrong.

(b) Again, this outcome seems very extreme, as only 1 value is above the hypothesised median. Again, the module team would conclude that the hypothesis is almost certainly wrong.

(c) This result is just what we would expect if the hypothesised median is equal to the true median. Thus the module team would conclude that the hypothesis is quite possibly true.

(d) With 7 values above and 5 below, we are only 1 value different from half above and half below. If we drew another sample, it could easily have 6 above and 6 below, or perhaps 5 above and 7 below. The module team would again conclude that the hypothesis is quite possibly true.

(e) Here you should be doubtful. The result is not very extreme, like cases (a) and (b), or very close to what we would expect, like cases (c) and (d). One could argue for (ii) that the hypothesis is probably wrong, though the module team would choose (i) that the hypothesis is quite possibly true. Some other phrase, such as ‘the hypothesis is a little unlikely’ would capture the view of the module team better.

Solution to Activity 28

(a) Since the outcomes are mutually exclusive we can use the addition rule to find the required probability. It is

$$\begin{aligned}&P(0[+]) + P(1[+]) + P(11[+]) + P(12[+])\\ &\quad\simeq 0.000 + 0.003 + 0.003 + 0.000\\ &\quad= 0.006.\end{aligned}$$

(b) The answer shows that the probability of obtaining one of these outcomes, when the population median is 0.98%, is extremely small. So if we observed one of these outcomes, it would seem sensible to reject the hypothesis and conclude that the population median is not equal to 0.98%.

Solution to Activity 29

(a)

$$\begin{aligned}P(0[+]) + P(1[+]) &\simeq 0.000 + 0.000\\ &= 0.000,\end{aligned}$$

$$\begin{aligned}P(0[+]) + P(1[+]) + P(2[+]) &\simeq 0.000 + 0.003\\ &= 0.003,\end{aligned}$$

$$\begin{aligned}P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) &\simeq 0.003 + 0.014\\ &= 0.017,\end{aligned}$$

$$\begin{aligned}P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) + P(4[+]) &\simeq 0.017 + 0.042\\ &= 0.059.\end{aligned}$$

(b) The probability of 0, 1, 2 or 3 values above the population median is the largest probability below 0.025.

(c) The 5% most extreme outcomes are $0[+]$, $1[+]$, $2[+]$, $3[+]$ and also $12[+]$, $13[+]$, $14[+]$ and $15[+]$. The last four are the same as $0[-]$, $1[-]$, $2[-]$ and $3[-]$. We should reject the hypothesis if there are three or fewer values on one side of the assumed median. The critical value at the 5% level for a sample of size 15 is 3.

(d) This corresponds to $1[+]$ and $14[-]$. The smaller value is 1. As 1 is less than 3, we should reject the hypothesis at the 5% significance level.

(e) The smaller value is 5. As 5 is greater than 3, we should not reject the hypothesis at the 5% significance level.

(f) The smaller value is 3. As this is equal to the critical value, we should reject the hypothesis at the 5% significance level.

Solution to Activity 30

(a) From Table 8, the critical value for a sample of size 21 is 5. So the hypothesis would be rejected at the 5% significance level if the outcome were any of $0[+]$, $1[+]$, $2[+]$, $3[+]$, $4[+]$, $5[+]$, $0[-]$, $1[-]$, $2[-]$, $3[-]$, $4[-]$ or $5[-]$.

(b) The critical value for a sample of size 32 is 9. So the hypothesis would be rejected at the 5% significance level if the sample contained nine or fewer values above the assumed median, or nine or fewer values below the assumed median.

(c) The critical value for a sample of size 7 is 0. This means that the hypothesis would be rejected only if the sample contained $0[+]$ or $0[-]$; in other words, it would be rejected only if all sample outcomes were on the same side of the assumed median.

Solution to Activity 31

(a) The median truancy rate for small schools in the East of England is 0.98%.

(b) Just 4 outcomes lie above the assumed median, and 19 outcomes lie below it. The test statistic is 4, the smaller of these numbers.

(c) From Table 8 (Subsection 4.1), the critical value for a sample of size 23 is 6.

(d) Since 4 is less than 6, we should reject the hypothesis at the 5% significance level and decide that the median truancy rate for small schools in the East of England is probably not equal to 0.98%. We can take our conclusion a little further. Since there are 19 values below 0.98% and only 4 values above it, we can conclude that the median truancy rate for small schools is probably *less than* 0.98%.

Solution to Activity 32

There are 28 outcomes (differences) and, if the median difference between the hybrids were 0 we would, on average, expect half the differences to be positive and half to be negative. In fact, only 7 of the differences are positive, while 21 of them are negative.

The smaller of these values is 7, so 7 is the value of the test statistic.

From Table 8 (Subsection 4.1), the critical value for a sample of size 28 is 8.

Hence the hypothesis is rejected at the 5% significance level. Hybrid B appears to give a higher yield than Hybrid A.

Solution to Activity 33

(a) The median truancy rate for schools in Essex is 0.98%.

(b) There are ten outcomes above the assumed median, and five outcomes below it. The test statistic is 5, the smaller of these numbers.

(c) From Figure 9, the p-value equals
$2 \times (0.000 + 0.000 + 0.003 + 0.014 + 0.042 + 0.092) = 2 \times 0.151 = 0.302.$

(d) The p-value is 0.302 (which is quite large), so there is little evidence against the hypothesis that the median truancy rate is 0.98% in schools in Essex.

Solution to Activity 34

(a) The smaller of 11 and 1 is 1, so we want

$$\begin{aligned}
&P(0[-]) + P(1[-]) + P(0[+]) + P(1[+])\\
&\quad = 2 \times [P(0[+]) + P(1[+])]\\
&\quad = 2 \times (0.000 + 0.003) = 0.006,
\end{aligned}$$

from Figure 6.

The p-value is 0.006, which is very small, so there is strong evidence that the median truancy rate for academies is not 0.98%. It seems likely that their truancy rate is higher than that.

(b) Now the test statistic is 3.

$$\begin{aligned}
&2 \times [P(0[+]) + P(1[+]) + P(2[+]) + P(3[+])]\\
&\quad = 2 \times (0.000 + 0.000 + 0.003 + 0.014) = 0.034,
\end{aligned}$$

from Figure 9.

The p-value is 0.034, so there is moderate evidence that the median truancy rate for secondary academies is not 0.98%. The sample data suggest that their truancy rate is lower than that.

Solution to Activity 35

(a) Four values are above 1.0, three values are equal to 1.0, and 16 values are below 1.0.

(b) The test statistic is 4 (the smaller of 4 and 16). Originally there were 23 observations. We discard the three observations that tied with 1.0, leaving a sample size for the test of 20.

(c) From Table 8 (Subsection 4.1), the critical value at the 5% significance level for a sample size of 20 is 5. As $4 < 5$, we reject the hypothesis at the 5%

significance level. The data provide moderate evidence that the median truancy rate of small schools in the East of England is not 1.0% – the median rate appears to be smaller than that.

Solution to Activity 36

(a) Nineteen values are above 19.7, three values are equal to 19.7, and eight values are below 19.7.

(b) The test statistic is 8 (the smaller of 19 and 8). Originally there were 30 observations. We discard the 3 observations that tied with 19.7, leaving a sample size for the test of 27.

(c) From Table 8 (Subsection 4.1), the critical value at the 5% significance level for a sample size of 27 is 7. As $8 > 7$, we do not reject the hypothesis at the 5% significance level. The data provide no clear evidence of climate change. (The p-value is actually 0.052, quite close to 5%, so there is some evidence that the stemplot data are not a random sample from the full set of data.)

Solutions to exercises

Solution to Exercise 1

This is a suggested ranking of the events, from the most likely to the least likely.

D. Death and taxes. (This gets top spot on the basis of the quote, ‘the only things certain in life are death and taxes’.)

F. You get exactly three heads when you toss a coin five times. (This has a 10-in-32 chance of occurring – as you will be able to calculate using the results in Section 3.)

A. A husband and wife find they were born on the same day of the week. (This is a one-in-seven chance.)

H. It snows in London on Christmas day. (Data from 1950–2006 suggests this is about a 6-in-100 chance.)

E. A mother’s first pregnancy results in twins. (Data for the United States suggest this is about a 3-in-100 chance.)

G. A chicken egg has two yolks. (The British Egg Information Service give this as less than 1-in-1000 chance.)

B. England will win the next football World Cup. (It’s probably not as likely as getting an egg with two yolks.)

I. You will be struck by lightning next year. (If you live in the United States, the probability is about 1 in 280 000.)

C. A Brazilian team will win the next football European Cup. (As Brazil is not in Europe, a Brazilian team cannot play in the European Cup.)

As in Activity 6, your order may be slightly different from this, but it should be fairly similar.

Solution to Exercise 2

(a) The probability that the man has blue eyes is

$$\frac{\text{number of men with blue eyes}}{\text{total number of men}} = \frac{2811}{6800} \simeq 0.413.$$

(b) The probability that the man has brown hair is

$$\frac{\text{number of men with brown hair}}{\text{total number of men}} = \frac{2632}{6800} \simeq 0.387.$$

(c) The probability that the man has blue eyes and brown hair is

$$\frac{\text{number of men with blue eyes and brown hair}}{\text{total number of men}} = \frac{807}{6800} \simeq 0.119.$$

(d) The categories ‘brown hair’ and ‘black hair’ are mutually exclusive, as a man cannot have both. Hence the probability that the man has brown hair or black hair is

$$\frac{\text{number of men with brown hair} + \text{number of men with black hair}}{\text{total number of men}}$$
$$= \frac{2632 + 1223}{6800} \simeq 0.567.$$

(e) The probability that the man does not have black hair is

$$\begin{aligned} &1 - \text{probability that he does have black hair}\\ &\quad = 1 - \frac{\text{number of men with black hair}}{\text{total number of men}}\\ &\quad = 1 - \frac{1223}{6800} \simeq 1 - 0.180 = 0.820. \end{aligned}$$

Solution to Exercise 3

If A is the event that Sue goes to the hockey match on Saturday and B is the event that her team wins, then A and B are statistically independent events (assuming that whether or not Sue does go to watch the match has no effect on how likely it is that her team wins), so

$$P(A \text{ and } B) = P(A) \times P(B) = 0.3 \times 0.4 = 0.12.$$

That is, the probability that Sue will watch her team play on Saturday and they will win is 0.12.

Solution to Exercise 4

(a) The number of ways of choosing three flags from seven flags in a specified order is

$$7 \times 6 \times 5 = 210.$$

Hence 210 different signals can be made.

(b) When a signal is flying there are four flags left in the box. The number of ways of choosing four flags from seven flags when order does not matter is

$${}^{7}C_{4} = \frac{7 \times 6 \times 5 \times 4}{4 \times 3 \times 2 \times 1} = \frac{840}{24} = 35.$$

Hence there are 35 different combinations of flag that could be left in the box when a signal is flying.

Solution to Exercise 5

The probability of obtaining a head is $\frac{1}{2}$. Hence

$$\begin{aligned} &P(\text{three heads in five tosses of a coin})\\ &\quad = {}^{5}C_{3} \times \left(\frac{1}{2}\right)^{5} = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} \times \frac{1}{32} = 10 \times \frac{1}{32} = 0.3125. \end{aligned}$$

(This is the same as P(three observations out of five are above the median).)

Solution to Exercise 6

The dealer's claim would be supported if the following were true.

The median petrol consumption of the car is 37 miles per gallon.

To determine the test statistic, we count the number of values above and below 37.0. There are 8 values above and 26 below. So the test statistic is 8.

From Table 8 (Subsection 4.1), the critical value for a sample of size 34 is 10. Since 8 is less than 10, we can reject the hypothesis at the 5% significance level and conclude that the median petrol consumption of the car is unlikely to be 37 miles per gallon. After looking at the sample, we can infer that the petrol consumption is probably less than 37 miles per gallon, and so the dealer's claim is not supported.

Solution to Exercise 7

There are 15 mice, and if the presence/absence of a mirror did not influence cage-preference, we would, on average, expect half the mice to spend more time in the cage with a mirror. In fact, only 3 of them did so, while 12 spent more time in the cage without a mirror.

The smaller of these values is 3, so 3 is the value of the test statistic.

From Table 8 (Subsection 4.1), the critical value for a sample of size 15 is 3.

Hence the hypothesis is rejected at the 5% significance level. We can conclude that mice appear to prefer a cage without a mirror.

Solution to Exercise 8

Three mice spent more time in the cage with the mirror, and 12 spent more time in the cage without a mirror.

The smaller of these values is 3, so 3 is the value of the test statistic.

From Figure 9, the p-value equals

$$2 \times (0.000 + 0.000 + 0.003 + 0.014) = 0.034.$$

This is moderately small – the p-value is between 0.01 and 0.05. Thus there is moderate evidence against the hypothesis. That is, there is moderate evidence that mice prefer a cage without a mirror.

Solution to Exercise 9

(a) Three values are above 0, four values are equal to 0, and nine values are below 0.

(b) The test statistic is 3 (the smaller of 3 and 9). Originally there were 16 observations. We discard the four observations that tied with 0, leaving a sample size for the test of 12.

(c) From Table 8 (Subsection 4.1), the critical value at the 5% significance level for a sample size of 12 is 2. As $3 > 2$, we do not reject the hypothesis at the 5% significance level. The data provide no clear evidence of methadone changing depression status.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover image: Minxlj/www.flickr.com/photos/minxlj/422472167/. This file is licensed under the Creative Commons Attribution-Non commercial-No Derivatives Licence http://creativecommons.org/licenses/by-nc-nd/3.0/

Introduction, cartoon (sharks and statistics): www.cartoonstock.com

Subsection 1.2 figure in Activity 3: Library of Congress

Subsection 1.2 cartoon (hula hoops): www.causeweb.org

Exercises on Section 1, cartoon (lightning): www.causeweb.org

Section 2 cartoon (fortune-teller): www.cartoonstock.com

Subsection 2.1 figure: ICMA Photos / This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Subsection 2.3, figure by Activity 15: Taken from http://cdn.foodbeast.com

Subsection 3.1 photo (Programme Committee): The International Statistics Institute

Subsection 3.1 photo (combinations of clothes): Taken from http://makingthingsbeautifulagain.blogspot.co.uk/2012/11/project-333-season-1-list.html

Subsection 4.1, figure by Activity 32: Alternative Heat / Flickr.com

Exercises on Section 4, figure by Exercise 7: http://about.me/rafaeltronquini

Subsection 5.2, figure by Activity 36: George Jansoone / This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Subsection 5.3 cartoon (data in perspective): www.causeweb.org

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index